

https://dspace.org/

Genome annotation of the 1.2MB Region on chromosome 8p22-p23.1 harbouring the gene for Keratolytic Winter Erythema (KWE)

Aron, Shaun Lyle

2012-01-17
http://hdl.handle.net/10539/11059

Downloaded from a DSpace repository. https://dspace.org/

Genome Annotation of the 1.2Mb Region on Chromosome 8p22-p23.1 Harbouring the Gene for Keratolytic Winter Erythema (KWE)



Shaun Lyle Aron

2011

A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfillment of the requirements for the degree of Master of Science in Medicine

Declaration

I, Shaun Lyle Aron, declare that this dissertation is my own unaided work. It is being submitted in fulfillment of the requirements for the degree of Master of Science in Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

_____

5th August 2011

Dedication

To my loving family and friends who have supported me

throughout my academic career thus far

Abstract

Keratolytic winter erythema (KWE) or Oudtshoorn skin disease is a rare autosomal dominant skin disorder for which the genetic cause remains unknown. The disorder manifests in the form of erythema and hyperkeratosis of the palmar-plantar regions and has been linked to a 1.2Mb region on chromosome 8p22-23.1 between markers D8S1759 and D8S552. A prevalence of 1/7200 has been observed in the South African Afrikaans-speaking white population with a lower unspecified prevalence occurring in the coloured South African population. A number of positional candidate genes within the critical region have been assessed for pathogenic mutations, however to date the causative gene has not been identified.

The objective of the current study was to examine the KWE critical region for highly conserved coding and non-coding regions and copy number variants (CNV) and to determine if these regions may play a role in the molecular etiology of the disease. Highly conserved regions were identified based on sequence conservation across a range of evolutionary diverse organisms. These regions were further analysed for possible protein-coding gene structure, regulatory motifs and RNA secondary structure. In addition, a custom CGH tiling array (384K Roche-Nimblegen) was used to identify CNVs across the extended KWE critical region in both affected and unaffected individuals. The multi-species sequence alignment revealed eight regions that showed a high level of conservation above a 70% threshold. Functional analysis of two of the conserved regions led to the identification of a novel protein-coding gene *deubiquitinating enzyme 3* (*DUB3*) within the critical region which presented as a credible functional candidate for KWE. Two of the conserved regions were identified within an open reading frame *c8orf13* which has previously been examined and found to contain no pathogenic mutations that segregate with the KWE phenotype. The remaining four highly conserved regions were found within non-coding sequence and computational analysis revealed putative regulatory motifs in the form of transcription factor binding sites.

The copy number variation analysis did not show evidence for the presence of any large or small consistent CNV alleles likely to impact on any of the functional candidate genes in the KWE critical region. No common CNV alleles were observed in all of the KWE affected individuals examined and showed absence in unaffected family members. A significant variation in copy number was however observed in affected individuals within a previously defined copy number variable *beta-defensin* gene cluster which has been associated with psoriasis. Although the exact copy number of the cluster could not be determined in the present study due to the cross hybridization between genes in the family, the CNV observed in affect individuals for the cluster suggests that it may be involved in the modulation of the clinical severity of KWE.

The present study has led to the identification of a previously uncharacterised novel gene *DUB3* within the KWE critical region which furthermore presented as a plausible functional candidate for the KWE phenotype. In addition, it has revealed that the molecular cause of KWE is unlikely to be exclusively due to copy number variation within the genes in the critical region. The current study has provided valuable insight into the KWE linked critical region and revealed a number of potential regions of interest to be examined in further studies exploring the molecular cause of the disease.

Acknowledgements

I would like to acknowledge the following persons:

Prof. Michèle Ramsay for her supervision, guidance, encouragement, support and advice

Prof. Scott Hazelhurst for his support and encouragement

Miss Angela Hobbs for her advice on the haplotype analysis and for her assistance with the laboratory investigations arising from the findings of this study

Dr. Peter Hull for providing funding for the copy number variation experiment

The National Bioinformatics Network (NBN) and National Research Foundation (NRF) for their financial assistance

Table of Contents

Declaration i

Dedication ii

Abstract iii

Acknowledgements v

Table of Contents vi

List of Figures xi

List of Tables xiii

Nomenclature xiv

Chapter 1. Introduction 1

1.1. Genomic approaches to identify disease genes for a monogenic disorder *4*

1.1.1 Identification of known genes 5

1.1.1.1 Candidate gene selection 7

1.1.2 Identification of novel genes and regulatory elements using comparative genomics 9

1.1.2.1 Highly Conserved Regions 10

1.1.2.2 Conservation vs Functionality 11

1.1.2.3 Assessing functionality of highly conserved regions 12

1.1.2.4 Transcription factor binding sites and promoter regions 13

1.1.2.5 Non-coding RNA genes 14

1.2 Copy number variation *15*

1.2.1 Identification of copy number variants 15

1.2.2 CNVs and disease 17

1.3 Gene regulation and disease *19*

1.3.1 Promoter mutations and transcriptional control of gene expression 19

1.3.2 RNA mediated gene regulation 22

1.4 Keratolytic Winter Erythema *26*

1.4.1 History of the disease 26

1.4.2 KWE Phenotype 27

1.4.3 The Skin 30

1.4.3.1 Maintenance of the epidermis 31

1.4.3.2 The Stratum Corneum 33

1.4.3.3 Palmoplantar Skin 34

1.4.4 Keratodermas 35

1.4.4.1 Palmoplantar keratodermas 36

1.4.5 Linkage analysis for KWE 38

1.4.5.1 KWE candidate genes 38

1.4.5.2 Complexities of the KWE critical region 40

1.5 Study rationale and aims *42*

Chapter 2. Subjects and Methods 44

2.1 Subjects *45*

2.2 Methods *46*

2.2.1 Assessment of KWE critical region and polymorphic inversion 46

2.2.2 Identification of highly conserved regions 46

2.2.2.1 Comparative genomic analysis of KWE region 47

2.2.2.2 Alignment 48

2.2.2.3 Visualization and identification of highly conserved sequences 50

2.2.3 In silico analysis of highly conserved regions 50

2.2.3.1 Prediction of gene structures 51

2.2.3.2 Identification of transcription factor binding sites 51

2.2.3.3 Identification of conserved RNA structures 52

2.2.4 Laboratory validation of identified functional regions 53

2.2.5 Comparative Genomic Hybridisation Array (aCGH) 53

2.2.5.1 DNA Extraction 53

2.2.5.2 Custom CGH tiling array design 54

2.2.5.3 Data Analysis of 385k tiling array 54

2.2.5.4 Visual Analysis using SignalMap software 55

2.2.5.5. Analysis using TM4 Microarray Software Tool- MultiExperiment Viewer (MeV) 56

2.2.5.6 Analysis of known genes 56

2.2.5.7 Identification of CNVs present only in affected samples 56

2.2.5.8 Assessment of potential functional impact of CNVs 57

Chapter 3. Results 58

3.1 Validation of KWE critical region 58

3.2 Comparative genomic analysis of KWE critical region 61

3.2.1 Identification of highly conserved regions 61

3.2.2 Assessment of highly conserved regions 64

3.2.3 Functional analysis of highly conserved regions 68

3.2.3.1 Prediction of gene structures 68

3.2.3.1.1 Gene structure in regions 1-3 71

3.2.3.1.2 Gene structure in regions 4 – 8 72

3.2.3.1.3 Further analysis of region 6-7 73

3.2.3.2 Prediction of transcription factor binding sites 75

3.2.3.3 RNA secondary structure prediction 78

3.2.3.3.1 RNA secondary structure in region 4, 5 and 8 78

3.3 Comparative genomic hybridization array of KWE critical region 82

3.3.1 Copy number variation within genes 82

3.3.2 MeV Analysis of samples for CNVs 84

3.3.3 Two regions of copy number variation for further investigation 88

Chapter 4. Discussion 89

4.1 Discussion outline 89

4.2 Comparative genomics for the identification of functional elements in the human genome 90

4.2.1 Identification of evolutionarily conserved elements 92

4.2.1.1 Conservation versus constraint 94

4.2.2 Identification of a novel gene *deubiquitinating enzyme 3* (*DUB3*) 95

4.2.2.1 Genome complexities of the DUB3 region 96

4.2.2.2 DUB3 as a possible KWE candidate gene 99

4.2.3 Transcription factor binding sites within conserved regions 101

4.2.3.1 Computational predictions of regulatory elements 104

4.2.4 Identification of RNA secondary structure 105

4.3 Copy number variation and KWE 107

4.3.1 CNV of the beta-defensin gene cluster 107

4.3.2 Copy number variation of known genes in the region 109

4.3.3 CNVs within potential regions of interest 110

4. 4 Limitations and future studies 112

4.4.1 Limitations 112

4.4.2 Future Studies 114

4.5 Conclusions 116

References 118

Online references 130

Appendices 131

Appendix A: Ethics Approval 131

Appendix B: GenomeVISTA Plots of KWE critical region 132

Appendix C: Coordinates of orthologous regions for each of the highly conserved sequences 138

Appendix D: Probe coverage of the custom Nimblegen-Roche CGH tiling array 142

Appendix E: Summary of the pairwise copy number variation analysis 143

Appendix F: CNV Pairwise Comparison Results 146

List of Figures

Figure 1.1 A partial screenshot of human chromosome 8:11 317 148 - 12 786 692 from the UCSC genome browser showing the position of the region with respect to the entire chromosome followed by horizontal tracks representing Refseq genes, mRNA transcripts, ESTs, a sequence conservation track and SNPs and repeats present in the region. Each item is clickable and provides a link to detailed information. (Accessed Dec 2010, NCBI36/HG18) 6

Figure 1.2 KWE patient's hands, showing severe hyperkeratosis and peeling of the epidermal layer and associated erythema with distinct margins. 29

Figure 1.3 KWE patient's feet following severe peeling. The outline of peeling regions can clearly be seen as well as erythema. 29

Figure 1.4 Diagrammatic representation of the compositon of the epidermis of normal skin showing position of the stratum lucidum only visible in palmoplantar skin (Segre, 2006). 32

Figure 1.5 Schematic diagram of the KWE critical region on 8p22-23.1 showing the location of the markers, REPD and REPP clusters and gaps. 41

Figure 2.1 Schematic flow diagram of the methodology employed in the present study. The two main aspects of the project included the identification of highly conserved regions within the KWE critical region using a computational approach and the identification of copy number variants within the KWE critical region using an aCGH approach in affected and unaffected individuals. 44

Figure 2.2 Pedigrees showing the ten individuals from the four families used in the CNV study together with the corresponding Nimblegen codes. 45

Figure 2.3 Approach to constructing multispecies sequence alignments with a human reference genome. The human sequence (center) is being aligned with two other sequences (above and below). The final alignment shows that nucleotides from the two other sequences do not need to retain their original order or orientation and they may be subjected to inversions (blue) or duplications (green) (Margulies et al., 2007). 49

Figure 3.1 Schematic representation of the position of the KWE critical region. (A) Shows part of chromosome 8p in the normal orientation, whilst (B) shows the altered position of the KWE critical region on the inverted chromosome based on the haplotype analysis. 59

Figure 3.2 A representation of the KWE critical region between markers D8S1759 and D8S552 showing RefSeq annotated genes in the region (UCSC Genome Browser – Accessed Jan 2010)..... 60

Figure 3.3 GenomeVISTA conservation plot showing the two KWE candidate genes *FDFT1* and *CTSB*. The blue peaks represent exons that show conservation of greater than 70%, while the pink regions

represent non-coding conserved regions. Each track is the alignment of the human genome sequence with (a) Chimp, (b) Rhesus, (c) Dog, (d) Mouse, (e) Chicken, (f) Fugu, (g) Zebrafish and (h) Frog..... 63

Figure 3.4 Genome VISTA conservation plot for the open reading frame c8orf13. The red boxes clearly show the conservation of the exons in the hypothetical open reading frame c8orf13. Each track is the alignment of the human genome sequence with (a) Chimp, (b) Rhesus, (c) Dog, (d) Mouse, (e) Chicken, (f) Fugu, (g) Zebrafish and (h) Frog..... 66

Figure 3.5 Genome VISTA plot showing conservation of a non-coding region (red box) upstream of a region (blue box) that showed two distinct regions of sequence conservation within a hypothetical gene which was previously uncharacterised. Each track is the alignment of the human genome sequence with (a) Chimp, (b) Rhesus, (c) Dog, (d) Mouse, (e) Chicken, (f) Fugu, (g) Zebrafish and (h) Frog 67

Figure 3.6 Minimum free energy predicted secondary structure for region 4..... 79

Figure 3.7 Minimum free energy predicted secondary structure for region 5..... 80

Figure 3.8 Minimum free energy predicted secondary structure for region 8..... 81

Figure 3.9 Part of tiling array showing the olfactory repeat regions REPD and REPP, gap positions and segmental duplications highlighted. The beta-defensin gene cluster within the olfactory repeat region is expanded below. 83

Figure 3.10 aCGH plot for the normalised data in two unaffected individuals for the region encompassing the *FAM90A* pseudogene cluster and beta-defensin gene cluster. Regions of reduced copy number can clearly be noted in the case of the green and red bars, with a distinct difference in copy number between the two samples being observed for the blue bar. 84

Figure 3.11 Two identified copy number variants and their locations in relation to KWE genes of interest..... 88

Figure 4.1 Regions on chromosome 8 surrounding DUB/USP17 family members. Diagrams of chromosome 8 (GRCh37 primary reference assembly) from bases (A) 7 185 000 – 7 205 000 (B) 7 820 000 – 7 840 000 and (C) 11 980 000 – 12 000 000. The approximate position of the identified DUB/USP17 sequences as well as the adjacent beta-defensin, FAM90A and olfactory receptor genes are indicated by the boxes illustrated in the key below (Burrows et al., 2010). 98

Figure 4.2 Diagrammatic representation of all the possible elements involved in the regulation of gene expression. Transcription factors can bind to specific sites that are either proximal or distal to a transcription start site. (Wasserman and Sandelin, 2004)..... 103

List of Tables

Table 1.1 Relative frequency of different types of mutations underlying disease phenotypes..... 8

Table 1.2 List of all known palmoplantar keratodermas, their mode of inheritance and associated genes 37

Table 2.1 Species used in comparative genome analysis 48

Table 3.1 Human chromosomal coordinates of regions showing high levels of conservation across at least 6 species with a cutoff of 70% over 100bp..... 65

Table 3.2 Gene structure prediction results showing the highest scoring gene predictions for each of the conserved regions using the two different prediction tools, GENESCAN and GENEID 69

Table 3.3 Transcription factor binding site predictions for conserved regions..... 76

Table 3.4 Regions identified as displaying copy number variation by visual analysis of normalized and 100bp averaged probe intensity plots in all 10 samples using SignalMap. These regions exhibited variation from the normal copy number and were either duplicated or deleted..... 83

Table 3.5 Copy number variation observed in affected samples in both the normalized and 100bp averaged datasets using MeV. 86

Nomenclature

A	Adenosine
aCGH	Comparative genomics hybridization array
BLAST	Basic Local Alignment Search Tool
Bp	Base pairs
C8orf13	Open reading frame 13 on chromosome 8
C	Cytosine
CCR5	C-C chemokine receptor 5
cM	centiMorgan
CNC	Conserved non-coding sequence
CNV	Copy number variant
CRM	*cis*-regulatory module
Cy3	Cyanine 3 dye
Cy5	Cyanine 5 dye
D8Sdi	Dinucleotide microsatellite marker on chromosome 8
D8Stet	Tetranucleotide microsatellite marker on chromosome 8
DEFB	Beta-defensin
DNA	Deoxyribonucleic acid
DUB3	Deubiquitinating enzyme 3
EDTA	Ethylenediaminetetraacetic acid
G	Guanine
GLI3	Zinc finger protein GLI3
HSP90	Heat shock protein 90

KWE	Keratolytic Winter Erythema
LCR	Low copy repeat region
miRNA	MicroRNA
ncRNA	Non-coding RNA
OMIM	Online Mendelian Inheritance in Man
ORF	Open reading frame
p	Short arm of chromosome
RefSeq	Reference sequence database
REPD	Distal repeat cluster
REPP	Proximal repeat cluster
RNA	Ribonucleic acid
SEG	Single exon gene
T	Thymine
TFBS	Transcription factor binding site
U	Uracil
UCSC	University of California, Santa Cruz
UTR	Untranslated regions

Chapter 1. Introduction

One of the underlying aims of genetic studies is being able to find sufficient evidence to link specific phenotypes with genotypes. The study of monogenic disorders has proven to be an indispensible avenue for elucidating gene function and understanding the normal physiological and pathological pathways in the human body. The human genome comprises approximately 25 000 well described protein-coding genes, a number which has been constantly changing in recent years due to our increased understanding of what features constitute a functional gene (Gerstein et al., 2007). Prior to the completion of the human genome sequence, variation within these protein coding genes were identified as the molecular cause of several Mendelian or monogenic disorders. Although the complete human genome sequence and methods such as genome-wide association studies and gene expression studies were unknown at that stage, linkage mapping and positional cloning approaches were used to link specific inherited phenotypes to chromosomal intervals and furthermore to mutations within genes in these critical regions. In conjunction with the analysis of family pedigrees, changes in the DNA sequence of a gene could be correlated with a disease and the effect of the change in DNA on the biological function of the encoded protein determined.

The completion of the human genome project and the availability of a human reference genome sequence provided much anticipation for the rapid discovery of the molecular basis of many more Mendelian disorders, however in the last 20 years variation in only 2912 genes have been found to be the underlying cause of simple Mendelian disorders (http://www.ncbi.nlm.nih.gov/Omim - Accessed December 2010). This represents a mere 22% of the repertoire of known protein coding genes, with the molecular basis of almost 1776 confirmed inherited monogenic disorders still remaining elusive (OMIM). This unexpected trend can be attributed to the focus of disease gene discovery in recent years moving away from the rare Mendelian disorders to the more common complex conditions such as asthma, diabetes, physiological disorders and cancers which are responsible for the

majority of human illness and mortality (Antonarakis and Beckmann, 2006). Although the identification of the molecular basis of these diseases is clinically significant, the study of monogenic phenotypes has provided much insight into the mutation process and understanding of many genetic phenomena and physiological processes. These genetic phenomena include uniparental disomy, parental imprinting and the discovery of multifactorial disorders involving the complex interaction of variation at multiple genetic loci and environmental factors.

Classical gene discovery methods for single-gene disorders are based on the assumption that the dissemination of a phenotype within a family is indicative of the transmission of a defect in a single gene (Badano and Katsanis, 2002). The localization of this specific gene is then identified through the process of successive linkage mapping in affected families to identify a small critical interval of usually less than 1cM on a specific chromosome followed by screening of candidate genes with functional relevance to the disease phenotype within this region for pathogenic mutations. Although a simplistic approach, the genes responsible for many Mendelian diseases have been identified using this methodology since in most instances Mendelian diseases are rare and the causative gene can easily be identified if enough affected families are studied. An important factor in this approach however is the ability to link the candidate genes within the critical region to the disease phenotype. In many instances genes within the critical region have no known functional relevance to the disease or contain no pathogenic variants that segregate exclusively with affected individuals. In this case the identification of the causative gene becomes more challenging to determine. The identification of causative genes underlying simple Mendelian disorders has provided great insight into providing more accurate diagnostic, prognostic and potential therapeutic tools for many of these disorders.

With the completion of the human genome project and the sequencing of the genomes of various other organisms our understanding of the complex structure and definition of a functional gene has constantly been under fundamental change. This has mainly been

attributed to the plethora of new laboratory techniques and bioinformatic tools being developed to analyse the large amount of experimental data being generated on a daily basis. Coupled with the discovery of complex patterns of dispersed regulatory elements within previously uncharacterized non-coding regions in the human genome, the operational definition of a functional gene has evolved significantly from simple "protein-coding regions" to a more complex definition. The implications for the identification of disease genes is therefore considerably altered since there appears to be additional functionally relevant regions within critical regions of the genome apart from the well characterized known protein-coding genes. With the availability of an extensive array of genomic sequences together with improved computational tools and advancements in technological platforms available for laboratory investigations, the opportunity now exists to extend conventional gene discovery methods to identify the molecular basis of many of the simple Mendelian disorders such as keratolytic winter erythema (KWE) for which the molecular basis remains unknown.

Keratolytic winter erythema or Oudtshoorn skin disease is a rare autosomal dominant condition for which the molecular etiology remains unknown. Although it has been linked to a critical region on the short arm of chromosome 8 by Starfield *et al*. in 1997 (Starfield et al., 1997), classical gene discovery methods have revealed no causative mutations within excellent positional candidate genes investigated within this region to date. Since no pathogenic mutations have been identified in the candidate genes within the KWE critical region it is possible that the molecular etiology of the disease may not be due to the presence of a deleterious variant within a protein coding gene that leads to a structural alteration of the protein's function, but is rather as a result of a quantitative change in gene expression. If this is the case, the causative mutation may not lie within a previously characterised protein-coding gene, but instead within an element in the critical region that is involved in the regulation of the genes in the region.

1.1. Genomic approaches to identify disease genes for a monogenic disorder

A vast majority of monogenic disorders occur as a result of a mutation in a single gene which encodes for a protein with altered functionality. Linkage mapping is the first step in the investigation of the molecular basis of monogenic disorders. It aims to localise the disease gene based on tracking the inheritance pattern of the disease within families using known polymorphic genetic markers, such as restriction fragment length polymorphisms (RFLP), microsatellites and more recently single nucleotide polymorphisms (SNP) which are known to vary between individuals. The positions of the genetic markers are known and allow for the observation of the segregation of a set of marker alleles within affected families and the determination of linkage to a critical chromosomal region based on a statistical process. In most cases, once a large critical region is defined, additional markers in the region are typed in order to reduce the size of the critical region by observing the inheritance of groups of markers or haplotypes and identifying rare informative recombination events that may lead to a further reduction in the size of the region. The critical region can then be defined as lying between specific markers for which the genomic location is known. Delineation of the critical region in monogenic disorders using linkage mapping alone has in previous studies been sufficiently accurate to identify disease genes, as was the case in the identification of the cystic fibrosis (Kerem et al., 1989) (Riordan et al., 1989), Fanconi anemia (Strathdee et al., 1992) and hemachromatosis (Feder et al., 1996) genes.

In some cases however linkage mapping provides a starting point to investigate further for possible candidate genes. Initial linkage mapping is then followed by linkage disequilibrium (LD) mapping to further narrow down the region in which the disease gene must lie. LD is a statistical method that depends on identity by descent using single markers. In most instances these markers are microsatellites and the allelic associations between these markers and the disease are compared between affected individuals and population matched controls. Greater power is attained in fine mapping of a region when groups of

markers or haplotypes are considered simultaneously. This approach has proven successful in the identification of regions of the genome associated with complex disorders.

Prior to the mapping of genes to the human genome the prioritisation of candidate genes was a laboratory intensive process which involved positional cloning to produce a physical map of the critical region followed by the identification of genes within the cloned DNA using methods such as exon trapping and CpG island mapping. These methods proved successful and led to the identification of disease genes in important disorders such as Duchenne muscular dystrophy (Koenig et al., 1988) and Marfan syndrome (Dietz et al., 1991). With the availability of the annotated human genome sequence these methods have been replaced by a variety of computational approaches for identifying and prioritising candidate genes for further investigation.

1.1.1. Identification of known genes

Following the completion of the human genome project all known protein coding genes have been mapped to the human genome and stored in online genomic databases such as RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) at the National Center for Biotechnological Information (NCBI) and EMBL-Bank at the European Molecular Biology Laboratory database (http://www.ebi.ac.uk/embl/). In addition, corresponding protein sequences and additional structural and functional information are stored in databases such as Swissprot/Uniprot (http://www.expasy.org/sprot) and Protein Information Resource (PIR) (http://pir.georgetown.edu) with integrated protein information stored in the Interpro database (http://www.ebi.ac.uk/interpro). These databases contain a multitude of sequence and annotation data, not exclusively for the human genome but in recent years have expanded to include a number of genomes from an array of species as they are being sequenced. All genes stored in these databases have been manually curated in some instances and experimentally validated and therefore provide a rich resource for all known genes within the human genome as well as associated annotation data such as genomic and

full length mRNA transcript sequences, expressed sequence tags (ESTs), known polymorphisms and protein structures.

Querying these large databases when searching for disease genes is not a trivial task, but is achieved through the use of online genome browsers such as Ensembl (http://www.ensembl.org/index.html) and the UCSC genome browser (http://genome.ucsc.edu/) which allow for the targeted retrieval of information from various databases using a specific query. Once the critical region for a disease has been determined by linkage analysis, a type of *in silico* positional cloning method can be used to identify all genes in the critical region using the genome browsers mentioned above. The critical region is usually defined by two flanking genetic markers such as microsatellites or SNPs and therefore all genes in the region can be obtained by using the genomic coordinates of the markers to view the critical region within the genome browser.

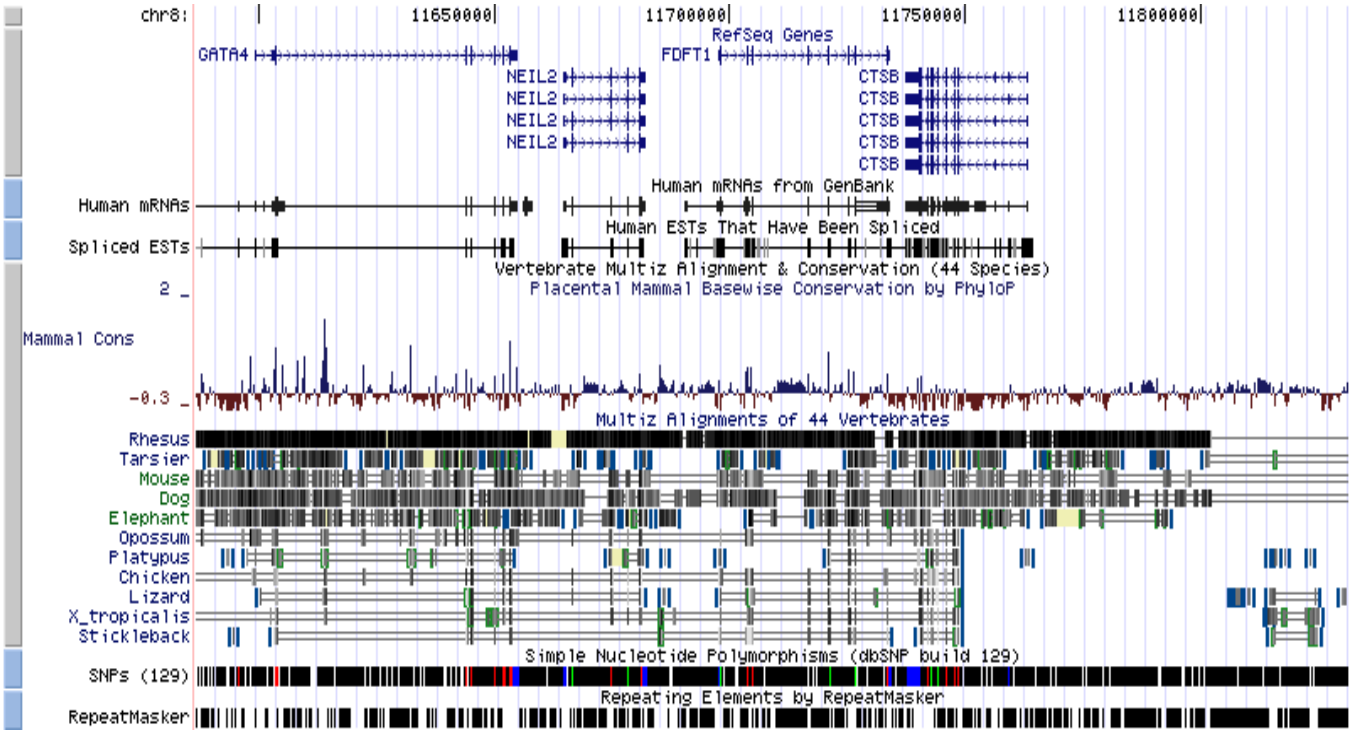


Figure 1.1 A partial screenshot of human chromosome 8:11 317 148 - 12 786 692 from the UCSC genome browser showing the position of the region with respect to the entire chromosome followed by horizontal tracks representing RefSeq genes, mRNA transcripts, ESTs, a sequence conservation track and SNPs and repeats present in the region. Each item is clickable and provides a link to detailed information. (Accessed Dec 2010, NCBI36/HG18)

1.1.1.1 Candidate gene selection

In theory every gene within the critical region could possibly be responsible for causing a disease, however prioritisation of the genes is required as laboratory screening of a large number of genes can become a costly and labour intensive process. Selection of candidate genes is based on linking the phenotypic manifestations of the disease to known biological functions of proteins encoded for by genes in the critical region. This is the starting point and clues to reduce the list of genes further may include analysis of all genes expressed in the main affected tissue in the disease to determine if any are encoded for in the critical region. In certain cases, a single strong biological candidate is identified and can be screened for mutations in affected individuals. In some instances however, a single strong candidate is difficult to determine and genes have to be prioritized for further analysis.

Although this approach has successfully been implemented to identify many disease associated genes, in some instances no promising candidate genes linked to the phenotype may be present in the critical region or not enough is known about the etiology of the disease to identify a functional candidate. In these cases although it may be more costly and labour intensive, an option is to screen all genes within the critical region for mutations if the number of genes is not too large. In both scenarios, screening of the genes may lead to no significant functional mutations being identified. At this point the exploration of novel genes or other functional elements in which the disease causing mutations may reside is the next option to identify the disease gene.

Current advancements in computational tools aimed at providing a deeper insight into the understanding and identification of all functional elements of the human genome have presented a vast new set of genomic elements that may be involved in monogenic disease pathology apart from mutations within well known protein coding genes. These genomic elements have been found to be involved with the direct regulation of known protein coding genes via a number of different mechanisms. With the abundance of genomic sequences

that have been generated in recent years, emphasis has been placed on developing efficient tools to annotate and characterize these unknown functional regions of the genome. Comparative sequence analysis of distantly related species has proven to be an extremely useful tool in the study and identification of evolutionarily functional regions in the human genome (Loots et al., 2000; Woolfe et al., 2005).

According to the Human Gene Mutation Database (HGMD) at the Institute of Medical Genetics in Cardiff (http://www.hgmd.cf.ac.uk/ac/ - Accessed December 2010) the majority of disease phenotypes identified to date are due to mutations within protein coding genes with a relative frequency of 56%; however regulatory elements account for just under 2 % (Table 1.1). Interestingly the relative frequency of mutations contributing to disease within regulatory elements is slightly greater than that of small indels (1.47%) and gross insertion/deletions (1.34%), while small deletions account for the second highest percentage of mutations at 16%.

Table 1.1 Relative frequency of different types of mutations underlying disease phenotypes

Mutation Type	Number	Percentage of Total
Missense/Nonsense	58984	56.10
Splicing	9969	9.48
Regulatory	1873	1.78
Small Deletion	16497	15.69
Small Insertion	6802	6.47
Small Indels	1545	1.47
Repeat Variations	328	0.31
Gross insertions/duplications	1414	1.34
Complex rearrangements	1014	0.96
Gross deletions	6709	6.38
Total	105135	

Data from the Human Gene Mutation Database – Accessed December 2010

1.1.2 Identification of novel genes and regulatory elements using comparative genomics

Comparative sequence analysis or the broad field of comparative genomics was born out of the development and success of high throughput sequencing technology. The ability to rapidly generate the complete DNA sequence of various organisms has spawned a movement away from the focus of traditional molecular genetics. We are now able to generate large amounts of data, and the emphasis has been placed on developing efficient ways to annotate and provide meaning to this data. Currently the complete or first assembly version for 56 vertebrate genome sequences is available via the Ensembl web browser (http://www.ensembl.org, Accessed December 2010). The systematic comparison of genomic sequences from different organisms forms the basis of the field that is comparative genomics. Comparative sequence analysis is based on the premise that functional regions of a genome are under selective pressure to remain unchanged and therefore should display a higher level of conservation across a wide range of species when compared to non-functional regions (Pennacchio and Rubin, 2001; Margulies et al., 2003). Comparative genomic approaches are now central to many areas of genomic research and exploratory studies.

Ultimately comparative sequence analysis aims to interpret and identify all functional regions within the human genome. The first comparative analysis was conducted with the human, mouse and rat genome sequences and it was found that approximately 5%-6% of the sequences were under selective constraints and therefore likely to have a functional role (Waterston et al., 2002; Gibbs et al., 2004). Of this 5%-6% only 1-2% is accounted for by protein-coding genes (Waterston et al., 2002). The remaining 4%-5% of conserved sequence corresponds to proposed functional non-coding regions, whose function remains to be elucidated. The comparison of orthologous genomic sequences has thus proven to be a powerful tool in the identification of both coding(Pennacchio and Rubin, 2001; DeSilva et al., 2002; Flicek et al., 2003) and non-coding (Dubchak et al., 2000; Wasserman et al., 2000; Kellis et al., 2003; Wasserman and Sandelin, 2004) functional elements in the genome. The non-coding regions of interest are those that are thought to be involved in the regulation of gene expression and therefore may be involved in disease (Hardison, 2000).

1.1.2.1 Highly Conserved Regions

The power of comparative sequence analysis is based on the ability to accurately and robustly align genomic sequences. Much effort has been placed into the development of algorithms to align long stretches of genomic sequences from a number of organisms and produce a meaningful visualization of the results. These algorithms aim to indentify conserved regions across multiple orthologous sequences in a manner that considers; a) the phylogenetic diversity of the originating species and therefore the general variation in sequence conservation, b) the varying neutral substitution rate for different genomic regions and c) the characteristics of the generated genomic multi-sequence alignment (Margulies and Green, 2003). The power of multi-species sequence alignments to identify highly conserved regions, as opposed to pairwise alignments, lies in the fact that the more evolutionarily distant organisms are, the lower the fraction of alignable sequence. Conserved regions identified in a multi-species sequence alignment with organisms as distantly related to humans as the fish and frog are therefore likely to be of more functional relevance than those regions conserved only in closely related mammals such as the mouse or chimp (Margulies et al., 2006).

Multi-species sequence alignments have therefore proven to be the most accurate and robust method by which to identify highly conserved regions in evolutionarily distant organisms that may be under active selective pressure to remain unchanged due to the presence of some unknown biological function (Margulies et al., 2006). Studies comparing the power of multiple sequence alignments to pairwise alignments on a defined genomic region revealed that a significantly lower percentage of highly conserved sequences were identified in humans when compared to pairwise alignments (Margulies et al., 2006). Many of these regions corresponded to known coding exons and UTRs. This suggests that more accurate and functionally relevant regions are indentified by multiple sequence alignments. Multi-species sequence alignment is a powerful tool for the identification of functional regions in the human genome whose discovery is difficult to elucidate from experimental methods (Hardison, 2000; Pennacchio and Rubin, 2001). With the availability of a large number of sequences from a diverse array of organisms there have been several recent

efforts to produce whole genome alignments across a range of species as well as develop efficient alignment algorithms and visualization tools to identify highly conserved regions.

1.1.2.2 Conservation versus Functionality

The vast majority of comparative studies have identified consistently high levels of conservation within coding exons of genes (Thomas et al., 2003). As much as 40% of the human genome can be aligned with the mouse genome and this percentage consists of both coding and non-coding sequences (Waterston et al., 2002; Schwartz et al., 2003). It is therefore unlikely that all of the non-coding homologous regions are conserved solely due to chance. It is widely accepted that non-coding regions displaying high levels of sequence conservation must be under selective pressure to remain the same due to functional constraints. As previously mentioned about 5% of the human genome is under purifying selection and therefore these regions must be involved in important functional processes.

The idea of conservation correlating to functionality is still under much debate as the constraint exerted on non-coding regions is difficult to quantify (Cooper et al., 2005). Previous studies have demonstrated that non-coding regions are under constraint and therefore purifying selection operates to ensure that they remain unchanged (Dermitzakis et al., 2003; Margulies et al., 2003; Cooper et al., 2005; Cooper and Brown, 2008). The challenge is determining the measure of constraint exerted on a specific non-coding sequence in order to have confidence in its functional relevance and filter out regions that should be considered as “comparative noise” (Thomas et al., 2003; Cooper and Brown, 2008).

1.1.2.3 Assessing functionality of highly conserved regions

Gene regulation was first described by Jacob and Monod (1961) in the lac operon of *E.coli* (Jacob and Monod, 1961). The operon consisted of a region of DNA termed the promoter, to which the RNA polymerase binds and a region called the operator to which regulatory genes bind (Jacob and Monod, 1961). Further studies revealed a range of other sequences which could affect every aspect of gene regulation, from transcription to mRNA degradation and post-translational modifications. These regions are found within coding sequences, flanking coding regions or may be far away from any known functional element (Gerstein et al., 2007). Regions found far away from known coding regions are known as enhancers, repressors or silencers and can directly affect the expression of distant genes due to the complex 3 dimensional folding of the chromatin structure. Regulatory elements therefore play an integral role in the expression of most genes, although they are not formally included in most of the definitions of a gene (Gerstein et al., 2007). Woolfe and colleagues showed that sequences within highly conserved regions are involved in vertebrate development in humans (Woolfe et al., 2005), while Nobrega *et al.* identified long range enhancers within non-coding gene desert regions of the human genome (Nobrega et al., 2003).

Non-coding regions that exhibit high levels of sequence conservation in humans across multiple sequence alignments are of great interest since this suggests that they are under active selective pressure to remain unchanged and are therefore likely to be of functional relevance (Pennacchio and Rubin, 2001; Boffelli et al., 2003; Ghanem et al., 2003). These regions are therefore potential novel genetics elements that may be involved in important biological functions but have not yet been characterised and experimentally validated. There are several possible regulatory elements that these highly conserved regions may encode for, including those involved in the spatial and temporal expression of known genes. These regulatory elements include conserved RNA structures, transcription factor binding sites, promoter regions, enhancers and silencers, CpG islands and unidentified ORFs. These regions, like known protein coding genes have the potential to harbor disease causing mutations.

In an effort to facilitate the characterization of highly conserved regions, several prediction algorithms have been developed to identify signatures associated with regulatory genomic elements. These algorithms aim to identify motifs and sequence signatures known to be associated with regulatory elements.

1.1.2.4 Transcription factor binding sites and promoter regions

The DNA consensus target sequences for many transcription factor binding sites are known and available in publically accessible databases such as TRANSFAC (Matys et al., 2003). These databases can be searched to identify potential transcription factor binding sites in a DNA sequence. Although most of these transcription factor binding sites are unique, many posses a small invariant core sequence surrounded by a number of degenerative nucleotides. This leads to a very high false positive discovery rate when searching the database with small individual sequences, although in recent years methods have been introduced to reduce this through the use of position-specific weight matrices (Matys et al., 2003). These database searches can reveal if there are motifs within a highly conserved sequence that is recognized by a specific transcription factor.

Promoter prediction databases such as the Eukaryotic Promoter Database, contain well annotated, RNA polymerase II promoters for which the transcription start sites have been determined experimentally (Perier et al., 2000). These sequences have been found immediately upstream of transcription start sites in known genes and can be searched to identify if a particular sequence shares common regions with known promoter regions. In contrast to the identification of transcription factor binding sites, it cannot be used to identify distant regulatory elements.

CpG islands are short stretches of DNA in which the frequency of CG nucleotides is higher than expected in the rest of the genome. Most of the Cs of the CG dinucleotide are methylated in the human genome. Unmethylated CpG islands are found at the 5’ end of several housekeeping genes and other regulated genes (Bird, 1986; Larsen et al., 1992). It has been shown that up to 56% of known human genes are associated with one or more CpG

island (Antequera and Bird, 1993). Methylated CpG islands are able to regulate expression since they have been linked to inactive genes, while unmethylated CpG islands promote gene expression. CpG islands may undergo selective pressure to remain unchanged in regions within close proximity of genes and in particular close to promoter regions. Software programs have been developed to identify CpG - rich DNA or CpG islands within genomic sequences.

1.1.2.5 Non-coding RNA genes

Non-coding RNA (ncRNA) genes are usually extremely difficult to identify, however their role in the regulation of genes via complex pathways has been well documented in recent years (Eddy, 2001; Mattick and Makunin, 2006). The first description of small, non-coding RNA mediated regulation of gene expression was observed in the *Caenorhabditis elegans* gene *lin14* which was found to be regulated by the *lin 4* gene product (Lee et al., 1993). Due to their lack of codons and no distinct ORF ncRNAs contain no distinct characteristics at the DNA sequence level. Comparative genomic analysis is therefore a powerful tool aimed at the identification of ncRNAs since these regions should be under selective pressure to remain unchanged (Frazer et al., 2003; Schwartz et al., 2003). Although no known sequence motifs are found within ncRNAS, the structural constraints of specific ncRNAs, such as characteristic pre-cursor secondary structure formation can be used to predict if a particular sequence is likely to be a ncRNA gene (Washietl et al., 2005). A program such as QRNA is able to identify possible RNA genes and is also able to detect regions of pre-miRNA, and mRNA predicted to have a conserved secondary structure, which may be involved in splicing, translational regulation, or mRNA localization or degradation (Rivas and Eddy, 2001).

1.2 Copy number variation

Comparative genomic approaches have revealed that the sequence of the human genome is thought to be about 99.9% identical among individuals in the world (Reich et al., 2002). This genetic variation occurs in the form of single nucleotide changes, variable repeat regions and microscopically visible chromosomal aberrations. In recent years an additional type of genetic variation in the form of copy number variants (CNVs) has been added to the list of phenomenon that differentiates us from each other at the DNA level (Sebat et al., 2004; Freeman et al., 2006; Redon et al., 2006; Sharp et al., 2006). CNVs are submicroscopic variations of DNA segments ranging from kilobases to megabases in size and include deletions, insertions, duplications and complex multi-site variants (Sebat et al., 2004; Sharp et al., 2006). These CNVs are found in all humans and can be as simple as tandem duplications or may include complex duplications or deletions of homologous sequences at several sites in the genome (Redon et al., 2006).

Basic human cytogenetics techniques were the original means by which to study and identify variation in whole chromosome copy number, rearrangements and other structural abnormalities. Microscopic chromosomal aberration could be efficiently identified and several chromosomal disorders were associated with these variants including Down's syndrome and various other congenital abnormalities (Kunze, 1980). The identification of submicroscopic CNVs was born out of the development of whole genome scanning techniques such as comparative genomic hybridization (CGH) arrays, which allow for the interrogation of the entire genome at a resolution higher than that of traditional cytogenetic techniques.

1.2.1 Identification of copy number variants

CGH is a classical cytogenetics technique used for comparing copy number of differentially labeled test and normal DNA using fluorescent *in situ* hybridization onto metaphase cells of a normal individual (Kallioniemi et al., 1996). Measurement of the fluorescence ratios across the entire chromosome gives an indication of the relative gain or loss in the test sample.

Array CGH (aCGH) is based on the same principle; however the differentially labeled test and normal DNA are hybridized together to arrays containing clones or synthesized probes covering the entire genome at a defined resolution or a specific region of interest.

Following the development of DNA array technologies several studies were conducted to analyse genome wide copy number variation. A genome-wide study identified 1 500 regions that showed variation in copy number, which included several genes and 360Mb of sequence (Redon et al., 2006). Another study revealed that several known disease-associated genes flank or fall within CNV regions with 58% overlapping with RefSeq genes and more than 99% overlapping with conserved non-coding sequences (Drake et al., 2006). A study conducted in a cohort of 60 patients with a history of pancreatic cancer identified 50 CNVs associated with the cancer (Lucito et al., 2007). A more focused study looked at copy number variation within 15 genes linked to schizophrenia; however no significant associations were identified (Sutrala et al., 2007).

CNVs, like other genetic variants, possess the potential to be benign or harmful depending on the region in which the variation occurs. CNVs can have a direct effect on the functioning of a nearby gene via at least five possible mechanisms and in some cases lead to disease. The first is dosage sensitive genes in which a CNV leads to the deletion or duplication of an entire gene. In addition, dosage insensitive genes can also be affected by CNVs if the deletion of one copy of a gene unmasks a recessive mutation on the homologous chromosome. Secondly dosage sensitive genes that overlap CNV regions can be disrupted by transversions, translocations or deletions resulting in reduced expression of that gene. Thirdly CNVs that are distant from dosage sensitive genes may occur within regulatory regions and alter expression the associated gene, or the inversion or translocation of a regulatory region may result in the aberrant expression of another gene. Lastly CNVs could function as susceptibility alleles, where the accumulation of multiple variants contributes to a disease phenotype (Feuk et al., 2006). Therefore CNVs can directly affect or alter gene expression through positional effects and lead to disease phenotypes.

1.2.2 CNVs and disease

With the advent of more efficient microarray technologies, the examination of the effect of CNVs in human disease has increased dramatically. In the field of cancer genetics the identification of CNVs within tumour cells has proven to be a powerful tool to predict tumour origins and aid in prognosis. This is attributed to the fact that in many tumour cells aberrations are known to affect tumour suppressor genes and oncogenes whose expression levels are altered by DNA copy number changes (Pollack et al., 2002; Pinkel and Albertson, 2005). The increased resolution of aCGH allows for the detection of deletions and duplications within tumour cells that were missed using classical cytogenetic techniques (Costa et al., 2008). A study conducted by Ruiz *et al*. in non small cell lung cancer (NSCLC) tumours using a combination of aCGH and expression arrays revealed a deletion of chromosome 14q23 together with reduced expression of *HSP90* which was encoded in the deleted region (Gallegos Ruiz et al., 2008). Interestingly however, the reduced expression of *HSP90* correlated with increased survival in affected individuals and *HSP90* inhibitors were found to have an anti-proliferative effect on tumour cells in vitro and led to the discovery of a novel drug target which has entered clinical testing (Gallegos Ruiz et al., 2008; Ruiz et al., 2008). de Wilde et al. showed that distinctive gains and losses observed from whole genome aCGH arrays could be used to accurately distinguish between positive and negative high risk human papiloma virus infected tumours in head and neck squamous cell carcinoma samples, aiding in the prediction of later stage progression of the tumours (de Wilde et al., 2008).

In addition to cancer studies, CNVs have begun to emerge as causal elements in simple Mendelian disorders. Gene deletions may not only be responsible for autosomal dominant conditions due to haploinsufficiency, but also cause autosomal recessive conditions in the presence of mutations inactivating the second allele. Balikova et al. identified one of the first Mendelian disorders to be caused by a tandem duplication of a 750kb region on chromosome 4 in individuals affected with an autosomal dominantly inherited syndrome characterized by microtia (deformity of the outer ear), eye coloboma (hole in one of the eye structures) and imperforation of the nasolacrimal duct (Balikova et al., 2008). The duplicated

amplicon encompasses a low copy region containing a cluster of olfactory receptor and defensin genes known to be variable in the studied population, however the exact amplicon of five tandem repeats was not identified in any control samples. Duplications can be responsible for monogenic disease by dosage effect when the gene and its regulatory sequences are duplicated in their entirety, but intragenic duplications can also lead to haploinsufficiency and to recessive disorders. This phenomenon was demonstrated by Thienpont *et al*. who identified a large intragenic duplication within the *ATRX* gene in ATR-X syndrome patients which resulted in a disruption of the transcription of the gene and the absence of ATRX mRNA and protein (Thienpont et al., 2007).

CNVs have been observed on chromosome 8p23.1 within the region of the beta-defensin gene family cluster containing *DEFB4, SPAG11, DEFB103, DEFB104, DEFB105, DEFB106 and DEFB107* (Hollox et al., 2003; Linzmeier and Ganz, 2005). This large repeat cluster shows copy numbers of between 2 and 12 in individuals in the UK (Hollox et al., 2003). Increased copy number of the beta-defensins within this repeat region has been linked to an increased susceptibility to the inflammatory skin disorder psoriasis in individuals of Dutch descent (Hollox et al., 2008). Another study found increased copies of the 360kb variable defensins region as well as expansion of the olfactory receptor repeat and duplication of the *GATA4* gene were linked to developmental delay and heart defects (Barber et al., 2005; Barber et al., 2008). The region of chromosome 8p23.1 is a structurally diverse region mostly due to the presence of repeat elements which makes the region prone to genomic rearrangements. Since the KWE critical region contains the proximal olfactory receptor repeat, it is possible that copy number variants may be present within this region and the surrounding genomic sequence.

1.3 Gene regulation and disease

Differential control of gene expression is one of the most important mechanisms to regulate gene function during development and overall homeostasis in the human body. Regulatory DNA regions act as switches to control whether genes are switched on or off in particular cell types and regulate gene expression levels to adapt to changing environments and physiological conditions (Strahle and Rastegar, 2008). Gene expression can be regulated at any of the seven possible control steps which include chromatin structure, initiation of transcription, processing of a transcript, transport to the cytoplasm, translation of mRNA, mRNA stability and protein activity stability (Villard, 2004). One of the main steps at which gene expression is regulated is initiation of transcription, which is mediated by *trans*-acting transcriptional activators or repressors (transcription factors) that bind to *cis*-acting regulatory recognition sites within close proximity or at a distance from a gene resulting in either an increase or decrease in transcription . Comparative genomic approaches have led to the realization that regulatory regions are usually highly conserved and many studies have shown that mutations in these regulatory regions can lead to dysregulation of linked gene products and in some cases may lead to disease. A genome-wide study by Stranger et al. examined the association of the transcriptional profiles of the 270 individuals in the four HapMap populations and 2.2 million common SNPs and identified 1 348 genes associated with *cis* regulatory signals (Stranger et al., 2007). The study revealed that SNP variants that lie close to genes, within regions usually involved in gene regulation, could possibly dramatically affect gene activity and be involved in disease.

1.3.1 Promoter mutations and transcriptional control of gene expression

Mutations in the human transcription factor *GLI3*, are known to cause a variety of developmental defect syndromes including Greig cephalopolysyndactyly syndrome (GCPS) (Kalff-Suske et al., 1999), Pallister-Hall syndrome (PHS) (Kang et al., 1997), postaxial polydactyly type A (PAPA) (Radhakrishna et al., 1997). *GLI3* is a key transcription factor during early development and is a primary transducer of Sonic hedgehog (*SHH*) signaling

which is responsible for multiple patterning steps during early development. A comparative analysis of the *GLI3* gene from the pufferfish (*Takifugu rubripes*) to human revealed 11 conserved non-coding regions within the intronic regions of the gene which are required for full scale transcriptional activation of the gene (Abbasi et al., 2007). Although expression of the gene was still activated when these conserved regions were deleted, the levels observed were not as high as the wildtype gene and indicated that these conserved regions were necessary for complete transcriptional activation of *GLI3* and may also harbor mutations that lead to the above mentioned syndromes. These 11 regulatory regions were observed within intronic regions of the gene; however it was shown that they directly regulate spatial and temporal expression of *GLI3*.

Ward *et al*. used a comparative genomics approach to indentify putative transcriptional regulatory elements in the non-coding region of the *Prop1* gene which is known to cause familial multiple pituitary hormone deficiencies (MPHD) (Ward et al., 2007). Although a majority of MPHD cases are caused by mutations in the protein coding regions of *Prop1* or an associated gene important for pituitary development, in some case no mutations have been detected in these genes. Three highly conserved putative regulatory non coding regions were identified by cross species sequence alignments of the extended gene sequence. Two of the regions were within the proximal promoter region of the gene within the first intron and another downstream of *Prop1* (Ward et al., 2007). All three regions were found to directly affect the spatial and temporal expression of the gene and therefore identified as an additional intronic region that should be screened for mutations in MPHD cases.

In addition to regulatory regions that lie within close proximity to the genes they regulate, regulators of gene expression may lie significantly far away from their gene targets. A combinatorial application of aCGH and the identification of evolutionarily conserved sequences was used to identify a long range regulator of the sonic hedgehog gene (*SHH*) (Klopocki et al., 2008). *SHH* plays an important role in defining the anterior-posterior axis of developing limbs. Point mutations in a region upstream to the *SHH* gene (termed ZRS) have been linked to triphalangeal thumb-polysyndactyly syndrome (TPT-PS), also called preaxial

polydactyly type II (PPD-II) which is a well defined autosomal dominant disorder that has been described in several families (Lettice et al., 2003). Further analysis of the region revealed that it was highly conserved in vertebrates with limbs and absent from those without limbs and was therefore characterized as a long range cis-regulatory element of *SHH* approximately 1Mb upstream of the gene within an intron of another gene *LMBR1* (Limb region 1 homolog) (Gurnett et al., 2007). A later study of a large family with variable TPT-PS which showed linkage to the same region revealed no mutations within ZRS and a whole human genome tiling path bacterial artificial chromosome (BAC) array of the region revealed a duplication of the region encompassing ZRS in affected individuals (Klopocki et al., 2008). This microduplication encompassing the regulator element ZRS leads to the same phenotype observed with the point mutations in the region and therefore is an additional mechanism by which long range regulation of genes may occur.

Glycosylphosphatidylinositol (GPI) deficiency is an autosomal recessive disorder which is caused by mutations within the *PIGA* gene. A point mutation 270bp upstream of the *PIGA* start codon was identified in two affected individuals and found to disrupt binding of the transcription factor SP1 to it promoter motif resulting in reduced transcription and GPI deficiency (Almeida et al., 2006). Familial hypercholesterolemia (FH) is an autosomal dominant disorder caused by mutations in the low density lipoprotein receptor gene (*LDLR*) which leads to a reduced clearance of plasma cholesterol from the body and increased risk of artherosclerosis and coronary heart disease. A vast majority of cases are caused by mutations within the coding region of the gene, however mutations have been observed 27bp and 60bp upstream of the transcription start site of the gene that have shown significant reduction of *LDLR* and only occur in affected individuals (Francova et al., 2004). Congenital erythropoietic porphyria (CEP), also known as Günther disease, is an autosomal recessive inborn error that results from the markedly deficient, but not absent, activity of the fourth enzyme in the hemebiosynthetic pathway, uroporphyrinogen III synthase (URO-synthase gene). Mutation analysis of the *URO-synthase* gene in CEP patients indentifies pathogenic mutations in 85-90% of cases; however in the remainder of cases only a single causative mutation has been identified (Desnick et al., 1998). Further analysis of the

sequence upstream of the transcription start site of the gene revealed four novel mutations clustered in a 20bp region that resulted in decreased expression of the *URO-synthase* gene and thus were identified as pathogenic mutations causing CEP (Solis et al., 2001).

In addition to disease causing mutations being identified in regulatory regions, mutations within regulatory regions have been found to be associated with altered susceptibility and progression of diseases. In human immunodeficiency virus-type1 (HIV-1) a variant within the promoter region of the *CCR5* gene has been associated with increased progression to AIDS in HIV affected cohorts (Martin et al., 1998). The *CCR5* gene encodes a cell surface chemokine receptor molecule that interacts with the CD4 co-receptor of macrophages and allows entry of the virus into the cell. The presence of this particular homozygous variant called CCR5P1 has been shown to be associated with faster progression to AIDS in three independent cohorts and therefore must, via some mechanism, result in an over expression of the CCR5 receptor causing an increased rate of infection of macrophages by the virus. To date a number of additional mutations in the *CCR5* promoter region of HIV-1 has been identified that alter progression of HIV to AIDS (McDermott et al., 1998; An et al., 2000; Knudsen et al., 2001).

1.3.2 RNA mediated gene regulation

Micro RNAs (miRNAs) are a class of small non-coding duplex RNAs 21-24 nucleotides in size which normally function as negative regulators of target mRNA expression at the posttranscriptional level. They regulate translation of target mRNAs by binding to the 3' untranslated region (UTR) resulting in target cleavage or translational inhibition (Ambros, 2004; Meister et al., 2004). miRNAs have also been shown to possess the functional ability to be positive regulators of gene expression in some instances (Vasudevan et al., 2007). The mechanisms by which miRNAs regulate gene expression are extremely complex, however a significant amount of information relating to miRNA processing and targeting and the phenomenon of RNA interference (RNAi) has been discovered since Fire and Mello and

colleagues identified the first double stranded RNA molecule responsible for post transcriptional gene silencing in *Caenorhabditis elegans* (Fire et al., 1998).

An abundance of functional miRNAs accounting for up to 5% of the human genome have been identified to date and act as key regulators in important biological processes such as cell growth, tissue differentiation, cell proliferation, embryonic development and apoptosis. Since the discovery of miRNA and the uncovering of their cellular mechanisms and functions, it has been shown that like conventional protein coding genes, mutations within these RNA genes can lead to various diseases via direct alteration of the expression of miRNAs and subsequently that of their gene targets. To date almost 230 diseases have been associated with miRNAs (http://cmbi.bjmu.edu.cn/hmdd, Accessed December 2010). miRNA gene silencing has therefore emerged as a key mechanism required for cellular maintenance and homeostasis. Like protein-coding genes and known regulatory regions, sequence variations within miRNA are likely to have a phenotypic effect and contribute to disease susceptibility if they perturb core components of the gene silencing machinery by altering the structure and expression levels of miRNA or target sites.

There exist two main classes of small RNAs that are involved in RNAi, small interfering RNAs siRNAs and miRNAs. In mammals, RNAi is mostly mediated by miRNAs which are 21-24 nucleotide duplexes incompletely base paired and form partial duplexes with the 3' UTR of targeted gene transcripts via an association with a complex structure of proteins termed the RNA-induced silencing complex (RISC) (Rana, 2007). miRNAs are the final product of a multistep process that begins with the generation of a primary miRNA transcript (pri-miRNA) which forms a hairpin structure containing one or more miRNA precursor. Their functional role is to inhibit mRNA translation and thereby regulate expression of their gene targets.

A number of miRNA genes are located in clusters within the human genome, with the largest being found on chromosomes 13q31.3 and 14q32.31. A miRNA cluster found on 13q31.3 has been associated with colorectal and lung cancer and the target is thought to be

an oncogene that enhances cell proliferation (Lanza et al., 2007; Matsubara et al., 2007). The largest known miRNA cluster is located at the imprinted Dlk1-Gtl2 domain on chromosome 14q32.31 containing 35 potentially expressed miRNAs within a miRNA containing gene (*Mirg)* (Tierling et al., 2006). This region has been found to be highly conserved between the human and the mouse genomes and contains a number of paternally imprinted genes which are expressed only from the maternal chromosome. The miRNAs encoded in this region are thought to play an important role in the regulation of imprinted genes within this region (Davis et al., 2005).

Although miRNAs are known to play an important role in gene regulation, the effect of mutations within these transcripts is still poorly understood. It is possible that polymorphisms which lie within regions important for formation of the pri-miRNA or target sequence may disrupt the function or processing of the miRNA. A study in which 173 known miRNAs were sequenced in 96 Japanese individuals indentified 10 SNPs in total. A single SNP mapped to a mature miRNA, *miR-30c-2*, which was likely to affect interaction of the miRNA with its target (Iwai and Naraba, 2005). Calin et al. identified a transition in the *miR-16-1* precursor in two individuals with chronic lymphocytic leukemia which resulted in reduced expression levels of *miR-16-1* in tumours. This SNP was not observed in 160 control subjects screened and reduced expression of the miRNA is thought to result in promoting cell proliferation (Calin and Croce, 2007). Many miRNAs have been found to be expressed at significantly lower levels in tumours, suggesting that they play an integral role in controlling cell proliferation and preventing tumorigenesis (Michael et al., 2003).

As previously mentioned, sequence variation within functional target sites of miRNAs may also lead to altered interaction with the miRNA. The first evidence that mutations in target sites can lead to an altered phenotype was observed in the study of Tourette's syndrome. The gene *SLITRK1* (Slit and Trk-like family member 1) was found to contain a frameshift mutation, which resulted in a truncated protein together with a G to A transition within the 3' UTR of the gene which was absent in healthy controls. Further analysis of the variant

revealed that it changed a G:U wobble match with a A:U match possibly facilitating the binding of *miR-189* and accentuating the downregulation of *SLITRK1.* The mutation within the target gene was responsible for stabilizing the interaction between *miR-189* and reducing expression of *SLITRK1*, thereby having a direct role in causing Tourette's syndrome (Abelson et al., 2005). Beetz *et al*. identified two point mutations in the 3' UTR of the gene encoding receptor expression –enhancing protein 1 (*REEP1*) in individuals suffering with hereditary spastic paraplegia. Mutations within *REEP1* have been observed in affected individuals. The 3' UTR mutations were observed within a highly conserved segment that showed exact complementarity to *miR-140*, known to be expressed in the central nervous system. One of the three mutations resulted in a change from a G:U wobble match to a A:U match possibly stabilizing the interaction between the miRNA and the target and therefore could result in downregulation of *REEP1* causing the disease (Beetz et al., 2008).

Several studies are now focused on the examining the role of miRNA genes in disease using high throughput arrays to analyse global miRNA expression in clinical samples, however as previously mentioned and noted in the examples above many of these functional miRNAs are evolutionarily conserved and can be identified using a comparative genomics approach. A number of these global miRNA expression studies are however providing insight into tissue specific miRNA profiles as well as identifying novel miRNAs. A recent study identified the first known skin-specific miRNA, *miR-203,* which was found to be expressed at the highest level in skin and over expressed in patients with the most common chronic inflammatory skin disease, psoriasis (Sonkoly et al., 2007). One of the targets of *miR-203,* cytokine signaling-3 (*SOCS-3*), which is a negative regulator of the STAT3-pathway which is activated by inflammatory cytokines and has important functions in the regulation of both innate and adaptive immunity as well as cell growth, survival and differentiation (Kubo et al., 2003). The over expression of *miR-203* leads to decreased *SOCS-3* levels and sustained activation of the STAT3 pathway resulting in prolonged skin inflammation in psoriasis patients. Although the over expression of *miR-203* has been observed in psoriasis patients, the driving force behind the increased expression is yet to be identified. In addition to *miR-203* a number of other

miRNAs exhibit altered expression levels in psoriasis and are implicated in the innate immune response (Sonkoly et al., 2007).

1.4 Keratolytic Winter Erythema

Keratolytic winter erythema (KWE) (OMIM 148370) is a rare autosomal dominant disorder which results in the epidermal keratinization and peeling of the skin of the palmar and plantar regions of affected individuals. Symptoms of the disease appear to be aggravated in the colder winter months as well as during seasonal changes. The disease was first observed and described by Findlay *et al.* (1977) in a group of South Africans residing in the small town of Oudtshoorn. The localized occurrence of this disease led to the belief that the disease was unique to South Africa and it initially became known as 'Oudtshoorn skin disease' or 'Oudtshoorn Vel'. Following the discovery and phenotypic characterization of the disease by Findlay, several cases were also observed in German families (Krahl et al., 1994; Starfield et al., 1997). Initially it was thought that KWE was unique to South Africa, however other cases were subsequently described in a large German family and an apparent spontaneous case in a 4 year old German child (Krahl et al., 1994).In South Africa, the prevalence of the disease is known to be 1/7200 in the white Afrikaans-speaking population, with a lower undetermined prevalence in the coloured population (Hull, 1986).

1.4.1 History of the disease

In the early 18th century a dermatologist in South Africa observed a number of individuals who exhibited similar, distinct clinical manifestations. The distinct nature of these symptoms eventually led to the idea that these individuals were afflicted with the same disorder. In 1977 Findlay led an inter-university study involving many dermatologists and medical professionals from various South African institutes to collect and assess information about the disease. The aims of the study included determining the general characteristics of the disease, naming the disease, identification of affected individuals, tracing of family history and establishment of genealogies. In addition, they hoped to carry out linkage studies using

serological polymorphisms and to establish a collection of information related to the disease including clinical data, photographs and histological material.

The study proved to be successful and led to the clinical name for the disease being coined as 'keratolytic winter erythema' as it emphasized the 3 main characteristics of the disease and 'Erythrokeratolysis hiemalis' being the alternative Latin equivalent. Although this was the clinical name decided upon, the geographical connection of the disease led to the additional names of 'Oudtshoorn skin disease' and 'Oudtshoorn vel' as previously mentioned. The linkage studies were not completed during this study however, based on the observed inheritance patterns in affected families, an autosomal dominant mode of inheritance was apparent. In addition, enough patients were seen in order to describe the vast range of clinical symptoms of the disease and the seasonal modulation of symptoms noted as a characterizing feature of the disease that allowed differentiation from other similar disorders (Hull, 1986).

1.4.2 KWE Phenotype

Clinical manifestations of KWE are characterized by initial reddening (erythema) of the palmar and plantar regions which occur in recurring cycles (Findlay et al., 1977). These lesions occur in a scattered arrangement and progressively enlarge with distinct red margins. KWE displays a distinct histopathological appearance which distinguishes it from other more common keratodermas. Although some individuals have complained of itching, in most instances there is no discomfort during the formation of these lesions. The progressive reddening is followed by thickening of the skin (hyperkeratosis) and leads to the formation of superficial dry blisters that originate from a central point from which peeling of the thickened skin develops (Starfield et al., 1997). The entire palmar and plantar region is usually affected with the dorsa being affected in some individuals. Multiple sites or lesions occur on the surface of both the hands and feet proceeding centrifugally until the entire upper layer of the skin becomes partially detached. Major skin creases have been found to form boundaries at which major peeling may be arrested (Hull, 1986), although in most cases the lesions do not extend past the edges of the palms and soles. Characteristic peeling is

noted in the web spaces of both the hands and feet. The peeling skin forms a thick scale which is not easily dislodged, leading to affected individuals removing the peeling layers to reveal a smooth new epidermal layer.

Perhaps the most interesting feature of this disease is the recurring cycles of the symptoms. Affected individuals exhibit seasonal variation in severity of symptoms, with the colder winter months aggravating the onset of symptoms, as well as seasonal changes. The colder drier weather appears to trigger increased thickening and peeling of the skin, whilst in some individuals, symptoms either subside or dissipate completely during the warmer summer months.

The clinical manifestations of KWE varies significantly between individuals and affected families, with symptoms ranging from mild localized peeling within webspaces, which is a distinct feature of KWE, to severe peeling and erythema of the entire palm or sole region (Hull, 1986; Starfield et al., 1997). Although colder winter months have been associated with an increase in symptoms in most individuals, this is not always the case and a few other factors have been associated with an increase or reduction of symptoms. These factors include palmoplantar perspiration and menstrual cycles in women which are associated with an increase in skin peeling and symptoms, while pregnancy in females and increased humidity levels found in coastal areas has been associated with a reduction or complete absence of symptoms (Hull, 1986; Starfield et al., 1997). The disease presents within the first 5 years and an early age of onset has been associated with increased severity.



Figure 1.2 KWE patient's hands, showing severe hyperkeratosis and peeling of the epidermal layer and associated erythema with distinct margins.



Figure 1.3 KWE patient's feet following severe peeling. The outline of peeling regions can clearly be seen as well as erythema.

1.4.3 The Skin

The skin forms the first line of defense against external physical and environmental factors that may pose a threat if allowed to freely enter the body. In addition to preventing organisms from entering, the epidermis also aids in the homeostatic regulation of internal conditions such as temperature regulation, retention of essential body fluids and the repair of damaged tissues caused by physical injury. The skin consists of three distinct layers, the epidermis, dermis and hypodermis which are distinct in appearance and origin. There exists a strictly controlled homeostatic balance between these three layers which results in continual development and maintenance of the skin throughout the lifetime of an individual (Ruth K. Freinkel, 2001).

The epidermis forms the outer protective covering of the skin and comprises a multilayered epithelium, the interfollicular epidermis and associated hair follicles and sebaceous and sweat glands. The epidermis is a resilient structure which is maintained throughout adult life by a continual production of stem cells found within the adult hair follicle, sebaceous glands and within the epidermis itself (Fuchs and Raghavan, 2002). The maintenance of this layer is important for the protection against infection and repair from physical or chemical damage that the skin is exposed to on a daily basis. The epidermis is composed of multiple cell layers resting on the basal membrane. The three main cell types are keratinocytes, melanocytes, Langerhans cells and Merkel cells. The main barrier of the epidermis is formed by the outer most layer which is consistently sloughed off and replaced with new cells derived from the inner layers. The maintenance of these cells is controlled by the process of terminal differentiation. Terminal differentiation is a process that results in the production of a specialized cell that can no longer divide (Niemann and Watt, 2002).

The epidermis is composed predominantly of keratinocytes, which make up about 90-95% of the cell population (Ruth K. Freinkel, 2001). These keratinocytes originate from the embryonic ectoderm during early development and are required for the production of keratins. Keratins form part of the structural component of keratinocytes and other epithelial cells and provide structural integrity together with microfilaments and microtubules.

Melanocytes are responsible for the production of melanin, a pigment responsible for the colour of skin, eyes and hair. Langerhans cells are primary cells responsible for the recognition, uptake processing and presentation of soluble antigens to the T-cells. They therefore function in a protective capacity by preventing unwanted substances from entering the epidermis (Hosoi et al., 1993). Merkel cells are believed to be derived from epidermal keratinocytes and are associated with nerve endings and therefore function as sensory mechanoreceptors in the skin (Gottschaldt and Vahle-Hinz, 1981).

The dermis is composed of the highest concentration of cells and contains the blood vessels, lymph vessels, sebaceous glands and hair follicles. The major cell type found in this layer is the fibroblast, which is involved in the synthesis and degradation of fibrous and non-fibrous connective tissue such as collagen and elastin (Ruth K. Freinkel, 2001). The hypodermis is the most important layer of the skin as it contains subcutaneous adipose fat which insulates the skin and is reinforced by collagen and elastic fibre. This layer binds the dermis to underlying muscles, stores lipids, insulates and cushions the body.

1.4.3.1 Maintenance of the epidermis

The epidermis is composed of five structurally defined layers; the basal layer, spinous layer, granular layer, stratum lucidum and the stratum corneum (Figure 1.4)(Segre, 2006). The stratum lucidum is only visible as a clearly distinct layer in palmoplantar skin. The maintenance of the epidermal layers is required in order to continuously provide a protective barrier against external harmful factors. This maintenance is achieved through the process of terminal differentiation of keratinocytes and the process of keratinisation (Niemann and Watt, 2002). Keratinisation is the process whereby rapidly dividing keratinocytes within the basal membrane withdraw from the cell cycle and break away from the basement membrane and travel towards the outer most layer of the epidermis. These cells undergoing a systematic process of differentiation from mitotically active basal cells to transcriptionally active spinous cells to enucleated granular cells as they travel towards the surface of the skin where they eventually differentiate into enucleated flattened squames in the stratum corneum (Barrandon and Green, 1987; Fuchs and Raghavan, 2002).

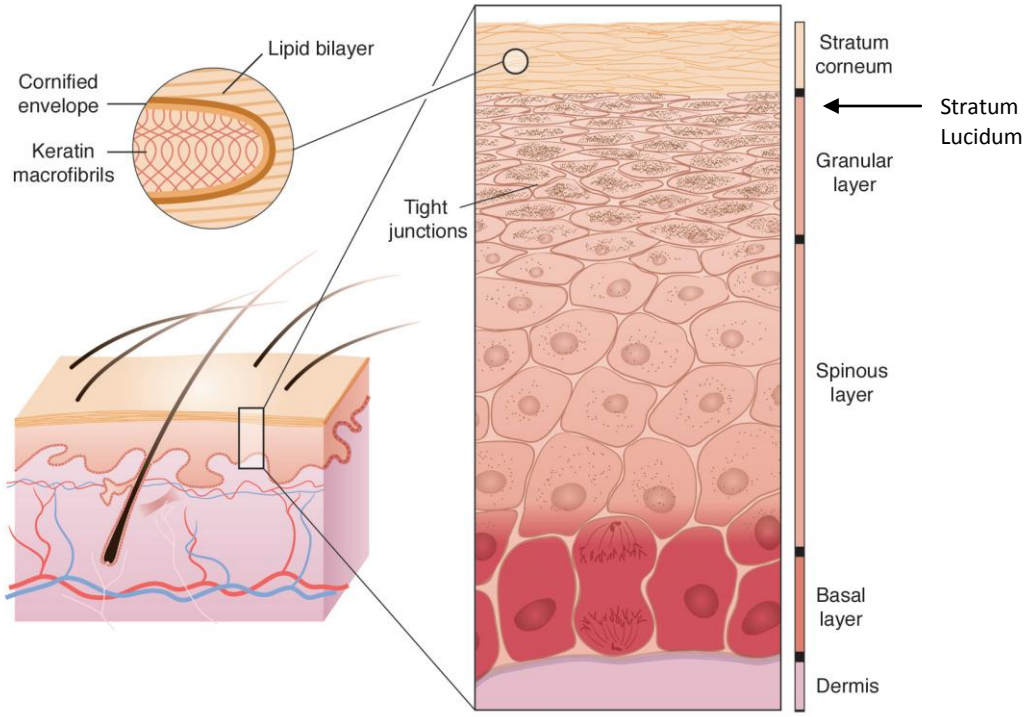


Figure 1.4 Diagrammatic representation of the compositon of the epidermis of normal skin showing position of the stratum lucidum only visible in palmoplantar skin (Segre, 2006).

The basal layer consists of basal cells which are attached to a basal membrane that is rich in extracellular matrix (ECM) proteins and growth factors (Fuchs, 2007). Basal cells attach to the basement membrane via cell-junction adhesion complexes composed mostly of intergrins . The continual division of the stem cells within the basal layer results in the constant production of cells which, once they commit to leaving the cell cycle, travel outwards towards the skin's surface.

Cells leaving the basal layer enter the spinous layer. As they enter this layer, the expressions of specific keratins are switched on (Fuchs, 2007). The spinous layer cells produce a reinforced durable cytoskeleton of keratin filaments which provide mechanical strength to the epidermis via large bundles of keratin filaments which mediate cell-cell connections known as desmosomes (Collin et al., 1992; Takahashi et al., 1995). The granular layer is composed of distinct keratinocytes which produce lipids within lamellar bodies and contain keratins bundled into macrofibrils together with filaggrin (Segre, 2006). The granular layer is

directly involved in the formation of the cornified envelope through the production of glutamine and lysine residues which are deposited under the plasma membrane of the cells. The cell membrane begins to degrade and becomes permeable to calcium which activates transglutaminase and results in the cross linking of the glutamine and lysine residues forming a reinforced sac around the keratin macrofibrils. This forms the scaffold of the cornified envelope after which the entire cell is destroyed and lipids are extruded and packaged in lamellar granules onto the cornified envelope. This is the final stage in terminal differentiation and leads to the formation of the stratum corneum which is made up of dead flattened cells (squames) composed of keratin macrofibrils and cross-linked cornified envelopes surrounded by lipids (Elias, 1983; Kalinin et al., 2002; Segre, 2006). The stratum corneum provides a permeability and antimicrobial seal that is replenished in a homeostatic manner as cells from the inner layers move outwards and are eventually sloughed off the skin's surface.

1.4.3.2 The Stratum Corneum

The stratum corneum varies in thickness in different regions of the body, as well as with age, sex and disease. The region of the palms and soles are composed of hundreds of layers, while regions such as the upper arm usually possess only 15 layers (Ruth K. Freinkel, 2001). The process of epidermal proliferation and differentiation or keratinisation is therefore a carefully balanced process, where too little proliferation results in thinning of the epidermis and loss of protection and too much leads to hyperproliferation. The homeostatic regulation of the stratum corneum in particular relies on the ability of the epidermis to determine when an increase in cell proliferation and migration is required, such as is the case when wounding of the skin occurs and under normal conditions to replenish cells that are continuously sloughed off the surface of the skin. Many keratodermas occur as a result of defects in the regulation of this important process and cause either hyperkeratosis as in the case of psoriasis and KWE or hypokeratosis observed in sun damaged or aged skin.

A majority of the protective function of the skin is provided by the stratum corneum (Elias, 2005). In addition to providing a physical barrier to pathogenic organisms, the epidermis also

produces an array of antimicrobial lipids, peptide toll-like receptors and chemokines that form part of the innate immune system (Gallo and Nizet, 2003; Ganz, 2003). The ability of the epidermis to control water loss and the entry of harmful microbes is due to the specific organization of the stratum corneum into two distinct regions. These two distinct regions are the lipid depleted corneocytes and the lipid enriched extracellular matrix. The molecular and chemical characteristics of the lipids in the extracellular matrix together with the structural integrity of the corneocyte account for the overall protective function of the stratum corneum (Elias and Feingold, 2001). Disruption or perturbation of the stratum corneum such as wounding results in a controlled cascade of events which leads to the proliferation of keratinocytes and T-lymphocytes which travel to the damaged skin in order to repair the wound and prevent infection. In addition the process of terminal differentiation is initiated to restore the epidermal barrier (Segre, 2006).

Defects in the process of terminal differentiation and maintenance of the stratum corneum lead to a number of skin disorders which manifest either in the inability to initiate terminal differentiation or defects which results in the uncontrollable proliferation of cells. The basis of the protective nature of the epidermis is due to the presence of keratinocytes which produce a system of dense keratin and keratin-associated proteins which provides the structural integrity of the epidermis. Disorders of keratinisation therefore occur as a result of mutations in any of the genes whose protein products are involved in this complex process.

1.4.3.3 Palmoplantar Skin

Although the process of terminal differentiation maintains the skin barrier throughout the surface of the body, regional variation in skin thickness and appearance does occur. There are two distinct types of human skin; hair bearing skin which covers most of the body surface and non-hair bearing or glaborous skin which is found on the palmoplantar regions. The composition and thickness of the epidermal layer of these two skin types varies significantly as well as the types of keratins that are expressed. The palmar and plantar epidermal layer has a significantly thick and compact stratum corneum, well defined stratum lucidium, has many sweat glands and a reduced number of melanocytes in comparison to the hair-bearing

skin (Swensson et al., 1998; Yamaguchi et al., 1999). Palmoplantar skin also has deep ridges which are believed to be an evolutionary specialization due to the amount of physical stress that has to be endured. Stem cells are located at the tips of these ridges and this is thought to occur since these elevated ridges are in contact with the most amount of stress and physical abrasion leading to a need for increased cell proliferation (Swensson et al., 1998). It was also observed that palmoplantar skin shows a more complex pattern of keratin expression with the expression of several keratins in addition to those expressed in hairy skin (Knapp et al., 1986; Moll et al., 1987; Langbein et al., 1994). Since keratins function as reinforcement molecules within keratinocytes, the increased number of keratins in palmoplantar skin may have evolved as a result of the extreme physical stress that it is exposed to. Keratins are therefore directly involved in controlling the thickening or keratinization of the skin and many genetic palmoplantar disorders have been linked to mutations in the various keratin genes.

1.4.4 Keratodermas

The first elucidation of the molecular basis of a human keratinizing disorder, epidermolysis bullosa simplex (EBS), was linked to keratin genes K5 and K14 following linkage of the disease to two keratin gene clusters (Bonifas et al., 1991; Coulombe et al., 1991). Analysis of the effect of the mutations on the keratin proteins together with the clinical phenotype of EBS revealed that these two keratins were involved in providing epidermal cells with mechanical strength (McLean and Irvine, 2007). In recent years several mutations in a number of keratin genes and other genes have linked to various genetic keratodermas. In most instances the disorders occur as a result of uncontrolled cell lysis in a specific region in which the mutated keratin gene is expressed. A recent review by McLean and Irving (McLean and Irvine, 2007) stated that 21 out of the 54 known keratin genes had been linked to monogenic disorders. In a large proportion of these disorders, the mutation leads to overproliferation of a specific layer of the skin, in most case the epidermis, together with a certain degree of overall fragility of the skin.

1.4.4.1 Palmoplantar keratodermas

Palmoplantar keratodermas (PPKs) are a diverse group of disorders which are associated with thickening of the skin on the palms and soles and have been linked to mutations that result in the excessive production of keratins. The overgrowth of the epidermis known as hyperkeratosis and associated erythema (reddening) are the most common symptoms observed in PPKs. In particular mutations in keratin 9, which is exclusively expressed in the palmoplantar epidermis leads to thickening and scaling of the palms and soles referred to as epidermolytic palmoplantar keratoderma (EPPK) (Reis et al., 1994). There are three clinical patterns of PPKs based on the type of epidermal involvement; diffuse, focal and punctuate. Both hereditary and acquired forms of PPK have been observed and there are various factors that contribute to the development of acquired PPKs, while many genes have been implicated as causative in inherited cases.

Palmoplantar keratodermas exhibit very similar clinical manifestations with a large overlap in symptoms and most clinical diagnoses require confirmation by histological examination of affected tissues (Itin and Fistarol, 2005). In recent years the molecular basis of several keratodermas has been identified and facilitates the accurate diagnosis of a particular disorder. Pathogenic mutations have been identified in genes that encode keratins, proteins involved in the development of the stratum corneum, cohesion molecules and transmembrane signal transduction enzymes (Kimyai-Asadi et al., 2002). The genetic basis of several PPKs has been identified to date (Table 1.2) and both autosomal dominant and recessive hereditary patterns have been observed. PPKs have also been observed as a symptom in the case of other complex disorders, such as Greither disease, which includes involvement of the Achilles tendon and hyperkeratosis on the knees and elbows (Fluckiger and Itin, 1993). The heterogeneity of PKKs has provided great motivation to identify the causative gene mutations in those PPKs for which they are not yet clear, to aid in the accuracy of clinical diagnoses.

Table 1.2 List of all known palmoplantar keratodermas, their mode of inheritance and associated genes

Type of Keratoderma	Mode of Inheritance	Gene
Vorner syndrome	AD	KRT9
Non-epidermolytic keratoderma	AD	KRT1
PPK with polycyclic psoriasiform plaques	AD	KRT1
PPK diffuse nonepidermolytic	AD	KRT1
Mal de Meleda	AR	SLURP1
Vohwinkel syndrome (classic)	AD/AR	GJB2
Locricrin keratoderma (Camisa)	AD	LORICRIN
Pachyonychia congentia I	AD	KRT6a, KRT16
Pachyonychia congentia II	AD	KRT6b, KRT17
Striate palmoplantar keratoderma	AD	KRT1, DSG1,DSP
Striate PPK/wooly hair/cardiopathy	AR	DSP
Naxos disease	AR	JUP
Papillon-Lefevre syndrome	AR	CSTC
Haim-Munk sydrome	AR	CSTC
Richner-Hanhart syndrome	AR	TAT
PPK with deakness	Mitochondrial	MTTS1
Tylosis with esophageal cancer	AD	EVPL
Epidermolysis bullbosa simplex	AD	KRT5
Epidermolysis bullbosa simplex	AD	KRT14
Ichthyosis hystrix Curth-Macklin	AD	KRT1
Ectodermal dysplasia/skin fragility syndrome	AR	PKP1
Hidrotic ectodermal dysplasia	AR	GJB6
Erythrokeratodermia variabilis	AD/AR	GJB4
Erythrokeratodermia variabilis	AD/AR	GJB3
KID sydrome	AD	GJB2
Vohwinkel syndrome	AD	GJB2
Bart-Pumphrey-Syndrome	AD	GJB2
PPK associated with sensorineural hearing loss	AD	GJB2
Darier disease	AD	ATP2A2
Dyskeratosis congenital	XL	DKC1
Dyskeratosis congenital	AD	TERC

AD= Autosomal Dominant AR = Autosomal Recessive XL = X-linked

(Itin and Fistarol, 2005)

1.4.5 Linkage analysis for KWE

KWE displays an autosomal dominant pattern of inheritance in affected families and the genetic cause of the disorder is unknown. Linkage studies using microsatellite markers distributed throughout the genome linked the disease phenotype to a 6cM region on chromosome 8p22-p23.1 between markers D8S550 and D8S552 (Starfield et al., 1997). Further studies using haplotype analysis reduced this region to a 1cM area between markers D8S550 and D8S265 based on an ancestral recombination event (Appel et al., 2002). Based on a genealogical study all KWE families could be linked back to a single founder from the late 18th century. This founder individual was a French sea captain named Francois Renier Duminy who worked for the Dutch East India Company and traveled to the Cape in 1775 (Hull, 1986). Haplotype analysis of KWE in South African families also presented further strong evidence for a founder effect in this population. The founder effect explains the relatively high frequency of the disease in the Afrikaans-speaking white and coloured populations of South Africa.

Appel *et al.*, produced a transcript map of the linked region and identified 5 genes in the KWE critical region, *BLK, MTMR8, TDH, AMAC1* and *C8orf13,* however sequencing revealed no mutations that segregated with KWE affected individuals (Appel et al., 2002). After the identification of an error in affection status in a critical cross-over family, the KWE locus was reassigned to a larger region on chromosome 8p22-23.1 between D8S1759 and D8S552. This critical region contains six candidate genes that have been identified and characterised. The genes include *FDFT1, CTSB, GATA4*, *NEIL2, BLK* and *LONRF1*.

1.4.5.1 KWE candidate genes

Screening for candidates within this newly defined critical region revealed two likely candidate genes, farnesyldiphosphate farnesyltransferase (*FDFT1)* and cathepsin B (*CTSB)* which are both involved in the maintenance of the skin epidermal layer. *FDFT1* encodes a membrane associated enzyme which plays a critical role in the mevalonate pathway responsible for the production of cholesterol, dolichol, ubiquinone, steroid hormones and

prenylated proteins in humans (Goldstein and Brown, 1990; Pandit et al., 2000). Cholesterol is an important intercellular lipid component found in the stratum corneum of the epidermal skin layer and is integral in the formation and maintenance of the permeability barrier together with an equimolar amount of free fatty acids and ceramides (Proksch et al., 1993). Additional studies have shown strong evidence that inhibition or increased levels of cholesterol results in alteration of the normal functioning of the epidermal barrier (Di Nardo et al., 1998; Kim et al., 2007).

CTSB belongs to the family of lysosomal cysteine proteases responsible for intracellular protein degradation (Schwartz et al., 2002). The cathepsin family of proteins has been found to be involved in a number of skin disorders involving defects in the homoeostatic maintenance of the epidermal layer. In addition, *CTSB* is thought to be involved in the regulation of keratinocyte migration via dissociation of cell-cell contacts of keratinocytes within the spinous layer (Buth et al., 2004). The migration of keratinocytes through the epidermis is essential for epidermal differentiation, keratinisation, wound healing and maintenance of the epidermal barrier.

These two genes were identified as excellent candidate genes for KWE and a fellow student, Angela Hobbs, has carried out an extensive mutation detection study of *FDFT1* and *CTSB* in KWE affected individuals, however no pathogenic mutations were identified within the exons and immediately flanking regions, or the mRNA transcripts. The variability in symptom severity and seasonal variation suggests that the molecular mechanism underlying the disease may be more complex than a conventional monogenic disorder characterised by a mutation in a protein coding gene. It is possible that the causal factor may be due to a quantitative change in gene expression due to a mutation in a regulatory region, the presence of a CNV within the region, or due to non-coding RNA or an RNA interference mechanism.

1.4.5.2 Complexities of the KWE critical region

The KWE critical region, lies within an area on the short arm of chromosome 8 that contains two low copy olfactory repeat clusters, a proximal repeat cluster (REPP) and distal repeat cluster (REPD) (Giglio et al., 2001). The REPD and REPP are 1.3Mb and 0.4Mb, however within these two repeats lie two sequence gaps of ~100 000bp for which there is no available sequence in the current build of the human genome (HG18) (Fig 1.5). The REPD is flanked by a cluster of beta-defensin genes that has been shown to display copy number variation (Barber et al., 2005; Hollox, 2008). These two repeat clusters are implicated in mediating a large inversion of approximately 3.5Mb that occurs within the region and which has also been observed in KWE affected individuals. The inversion has been observed at polymorphic frequencies (Giglio et al., 2001; Shimokawa et al., 2004).

Although the presence of the inversion is not directly harmful, it has been found to interfere with the correct pairing of chromosomes and increases the chance of non-allelic crossovers between the olfactory receptor gene repeats (Giglio et al., 2001). This event leads to additional chromosome rearrangements such as inverted duplicated deletion 8, supernumerary chromosomes with functional neocentromeres outside the normal centromere domain (analphoid chromosomes) and translocations in the offspring of heterozygous carriers of the inversion (Giglio et al., 2001; Shimokawa et al., 2004). These rearrangements result in genetic disorders associated with mental retardation, heart defects and facial abnormalities in affected individuals (Giorda et al., 2007).

The inverted form of the chromosome has been observed in KWE affected individuals but the polymorphic frequency of its occurrence suggests that it is a harmless genomic rearrangement. The breakpoints of the inversion lie within the REPP and REPD, however their exact locations are still unknown. The approximate breakpoints of the inversion have however recently been documented (Giorda et al., 2007). The KWE critical region lies within the inverted region of the short arm of chromosome 8 between markers D8S1759 and D8S552. KWE was initially linked to this region, which contains 6 known genes, however in the presence of the inversion, the position of D8S1759 could be altered and the region could

possibly encompass a much larger area. This is a significantly larger region and a fellow student typed further markers to identify haplotypes in families with the critical cross-over events in order to redefine the KWE critical region.

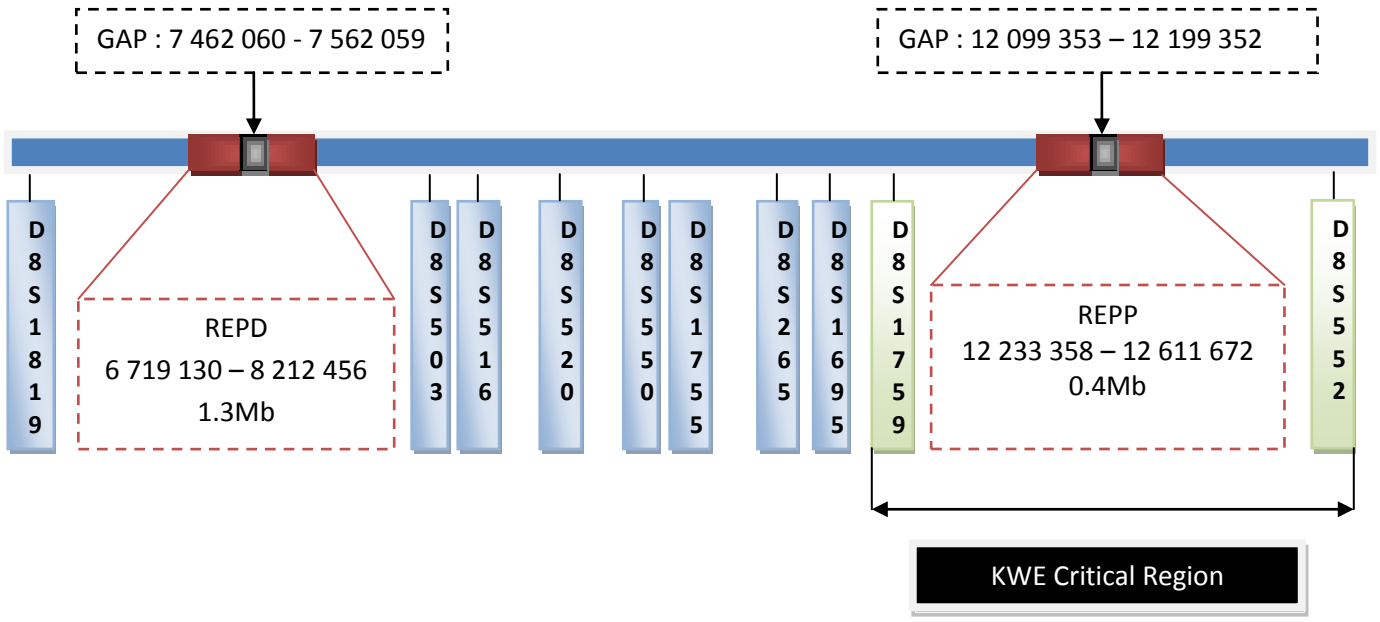


Figure 1.5 Schematic diagram of the KWE critical region on 8p22-23.1 showing the location of the markers, REPD and REPP clusters and gaps.

The presence of repeats and the inversion in the KWE critical region has made it a very challenging task to identify the genetic cause of the disease. The linkage of KWE to this region and distinct autosomal dominant inheritance confirm that there is a monogenic cause for the disease, however no mutations have to date been identified in the known positional candidate genes identified in the critical region. The unusual seasonal variation of symptoms and varied penetrance of the disease suggests that the molecular etiology of the disease may involve a quantitative aspect of gene expression, rather than a structural variant of a protein. If this is the case, the genetic cause of KWE may lie in an unknown functional region of sequence involved in regulatory gene expression that is yet to be identified and characterized.

1.5 Study rationale and aims

As highlighted in the above literature review, classical methods used to identify genes involved in simple genetics disorders have undergone significant changes in recent years with the discovery of previously undefined functional elements within the human genome. Many of these previously undefined regions lie within non-coding regions of the genome and play an integral role as gene regulatory elements. Furthermore it has been found that regulatory elements found in non-coding regions are highly conserved across different species. These elements have been overlooked in past investigations to identify genes involved in Mendelian disorders and may be the underlying cause of single gene disorders for which the molecular cause remains unknown.

Secondly, it has been observed that in addition to the presence of genetic variation in the form of microscopically visible chromosomal rearrangements and single nucleotide changes, there is an abundance of copy number variation of DNA segments throughout the genome. Previous studies have shown that like single base changes within protein-coding genes in the genome, these CNVs possess the ability to affect phenotype and cause disease.

As previously mentioned, the molecular basis of KWE is yet to be identified. Much work has gone into linking the disease to a candidate region and there is sufficient evidence to suggest that the disease is inherited in an autosomal dominant pattern and has a strong genetic component. Studies of candidate genes in the region have yielded no significant results with respect to the presence of pathogenic mutations.

Thus, the objective of this study was to examine the KWE critical region on 8p22-23.1 for highly conserved coding and non-coding sequences and regions of copy number variation and to further characterize these regions to determine if they may play a role in the molecular etiology of the disease.

This objective was achieved via two main aims:

Aims:

1. To search and identify highly conserved regions within the KWE critical region and further annotate and characterize these regions to determine if the may play a role in KWE.

2. To examine the KWE critical region for copy number variation and assess the impact of the variation on the region and the KWE phenotype.

Chapter 2. Subjects and Methods

Two approaches where employed in this study to annotate and examine the KWE linked critical region and are summarized in Figure 2.1.

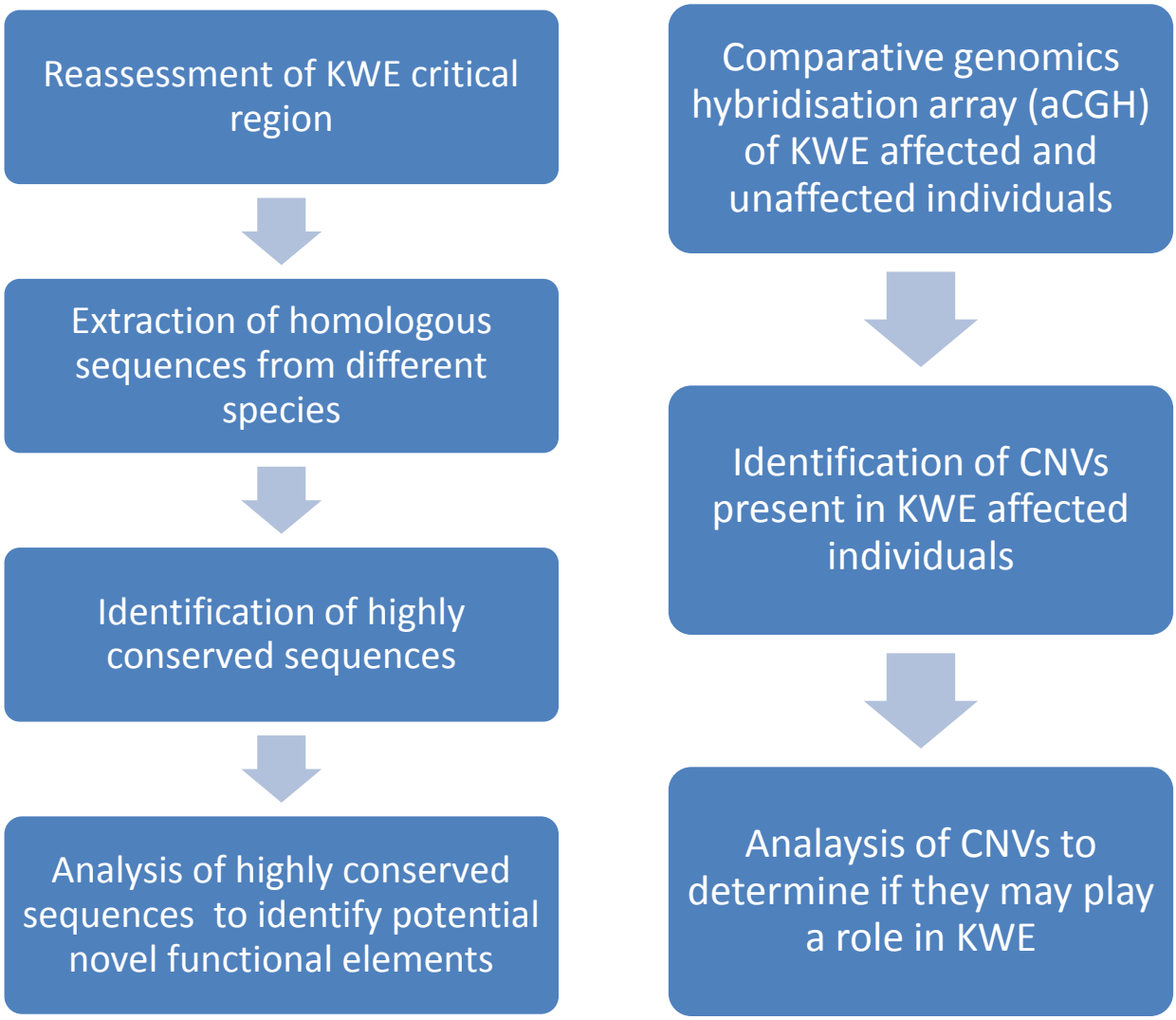


Figure 2.1 Schematic flow diagram of the methodology employed in the present study. The two main aspects of the project included the identification of highly conserved regions within the KWE critical region using a computational approach and the identification of copy number variants within the KWE critical region using an aCGH approach in affected and unaffected individuals.

2.1 Subjects

DNA samples were obtained from a cohort of South African KWE families of which four families were selected for the CGH fine tiling array experiment representing a total of 10 samples. Three of these families were composed of an affected and unaffected individual, while the remaining family consisted of four related individuals as indicated in the pedigrees (Figure2.2). A total of 5 KWE affected and 5 unaffected samples were used and following informed consent, 10ml of blood was collected in EDTA tubes from each individual. The samples for this project were collected as part of a concurrent KWE study conducted by a colleague, Miss Angela Hobbs, and ethics clearance for the use of the samples was obtained from the Human Research Ethics Committee, ethics approval number M070423.

Nimblegen No.	KWE No.	Affection Status
7732502	KWE46_1.1	Affected
7739102	KWE46_1.2	Unaffected
7732402	KWE46_2.1	Unaffected
7739302	KWE46_2.2	Affected
7737602	KWE23_1.2	Unaffected
7755202	KWE23_2.1	Affected
7737902	KWE45_1.2	Unaffected
7738002	KWE45_2.2	Affected
7739402	KWE48_1.2	Affected
7738902	KWE48_2.1	Unaffected

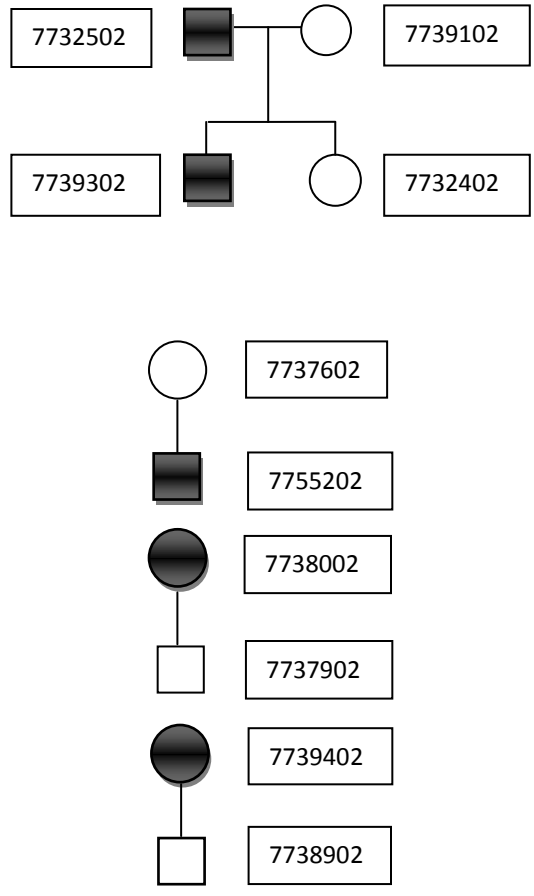


Figure 2.2 Pedigrees showing the ten individuals from the four families used in the CNV study together with the corresponding Nimblegen codes.

2.2 Methods

2.2.1 Assessment of KWE critical region and polymorphic inversion

A detailed analysis of all previous data on the polymorphic inversion on 8p22-23.1 was reviewed to assess the impact of the inversion on the KWE critical region. Although the presence of the inversion was known prior to the identification of the KWE critical region, the exact breakpoints were not well defined and were not considered in previous KWE studies. The positions of the approximate proximal and distal breakpoints of the inversion were reassessed in order to determine if they directly affect the KWE critical region. In addition, a previous KWE haplotype study (Starfield et al., 1997) was reviewed in order to confirm the correct order of markers within the inverted region. Using the data pertaining to the inversion together with the current marker order for the region obtained from the UCSC genome browser (http://www.genome.ucsc.edu/) a physical map of the KWE region was generated. Two genetic maps of the extended KWE critical region were constructed; one representing the order of markers on the inverted chromosome and the other representing the non-inverted chromosome. Based on these genetic maps a colleague, Miss Angela Hobbs, retyped markers within the KWE region and those flanking the breakpoints in KWE families in order to confirm the KWE critical region and identify markers that segregated with affected individuals. This information was used to define the KWE critical region to be assessed further in this study.

2.2.2 Identification of highly conserved regions

The human genomic sequence for the KWE critical region was obtained from the UCSC genome browser from approximately 200 000bp upstream of marker D8S1759 to 100 000bp downstream of marker D8S552. The human genomic sequence for the KWE critical region therefore corresponds to the March 2006 assembly (HG18), i.e. chromosome 8:11 317 148 – 12 899 353. For all further analyses, the coordinates of the regions mentioned correspond to

those in the human reference genome sequence. The identification of evolutionarily conserved genomic sequences is a potentially powerful means by which to identify functional regions of the genome. Conserved non-coding regions (CNCs) among mammalian and more distantly related organisms provide further regions of the genome that may harbor genetic variants that contribute to disease in addition to known protein-coding regions. In this study, the KWE critical region was examined to identify evolutionary conserved regions using a computational approach which facilitates the comparison and visualization of genomes of various species.

2.2.2.1 Comparative genomic analysis of KWE region

The choice of species was based on the ability to extract enough biological information from the multispecies alignment as possible. The ideal pairwise comparison to identify novel regions of functional relevance is based on the comparison of genomes from two physiologically or biologically similar organisms. However, there must be a balance between biological relevance, evolutionary distance and sequence analysis in order to identify sequences that appear to be under evolutionary constraint to remain unchanged in the background of sequence that has diverged due to random genetic drift. Multispecies sequences alignments of divergent species allows for a balance in the amount of sequence conservation observed as the comparison of closely related species results in an abundance of conservation obscuring functional regions, while the comparison of divergent species shows too little conservation and are less informative. To achieve this 8 species with a range of evolutionary distance from humans were chosen to conduct the multispecies sequence alignment and are listed in table 2.1 together with the corresponding genome builds.

Table 2.1 Species used in comparative genome analysis

Species	Common Name	Genome Build
Pan troglodytes	Chimpanzee	March 2006
Macaca mulatta	Rhesus Macaque	Jan 2006
Canis lupus familiaris	Dog	May 2005
Mus musculus	Mouse	July 2007
Gallus gallus	Chicken	May 2006
Takifugu rubripes	Puffer Fish (Fugu)	v.4.0.
Danio rerio	Zebrafish	March 2006
Xenopus tropicalis	Frog	v.4.1.

2.2.2.2Alignment

The first step in comparing genomic sequences is to map the letters of each DNA sequence to each other. This is achieved through the computation of an alignment which takes into consideration the selection of a reference or base genome against which sequences are compared on a pairwise basis. The human KWE critical region was selected as the reference sequence and all other sequences where then aligned to it on pairwise basis. In this study a combination of a global and local ("glocal") alignment technique was used to produce a multiple sequence alignment of the KWE critical region across 8 species.

The selection of a glocal alignment method was based on previously noted shortcomings of both local and global alignments to identify highly scoring local alignments within a larger syntenic region that has undergone a rearrangement event. GenomeVISTA , one of the tools available from the VISTA suite (http://genome.lbl.gov/vista) utilizes a modified version of a local alignment algorithm, Shuffle-LAGAN (S-LAGAN) to produce pairwise alignments of the human genomic sequence and another species using a local aligner such as BLAT to indentify

orthologous sequences and then progressively extends the pairwise alignments using a species phylogenetic tree to produce an accurate multispecies sequence alignment. The species tree is used to decide at what stage syntenic blocks of sequence should be aligned and is important since species with greater sequence similarity should be aligned first to obtain an accurate multi-species alignment. Glocal alignments therefore ensure that smaller high scoring local alignments are properly captured within the context of the global alignment.

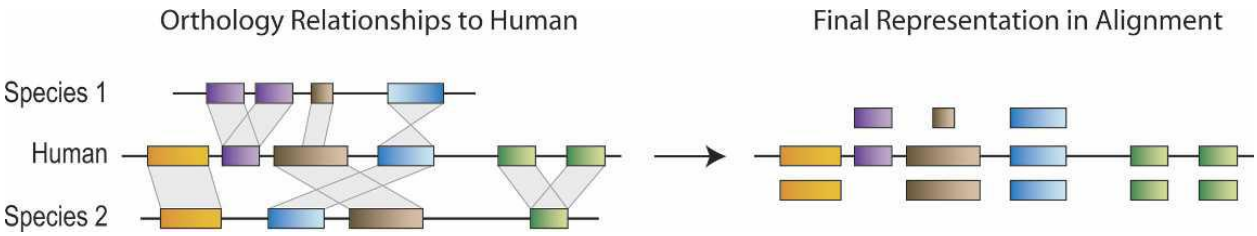


Figure 2.3 Approach to constructing multispecies sequence alignments with a human reference genome. The human sequence (center) is being aligned with two other sequences (above and below). The final alignment shows that nucleotides from the two other sequences do not need to retain their original order or orientation and they may be subjected to inversions (blue) or duplications (green) (Margulies et al., 2007).

The human genomic sequence for the KWE critical region of approximately 1.6Mb was submitted to the VISTA server in partitions of 300kb. The region was divided into 300kb intervals in order to efficiently view and explore the resulting alignments within the VISTA browser and extract sequence data for further investigations. GenomeVISTA allows for the selection of various parameters to be used when viewing the multi-species sequence alignment and includes the window size, which defines the smallest conserved region to be considered and the percentage conservation level. These parameters do not have any effect on the pairwise alignments but define the regions of interest observed when viewing the final multispecies sequence alignment. The multispecies alignment was visualized using a sliding window of 100bp and a minimal conservation score of 70% identity across all 8 species in order to identify regions of 100bp or larger that showed high levels of sequence conservation within the KWE critical region. These criteria were chosen on the basis that

many studies have shown that regulatory elements showing experimentally validated functional properties were found to meet these criteria (Loots et al., 2000; Donfack et al., 2005; van Deursen et al., 2007).

2.2.2.3 Visualization and identification of highly conserved sequences

The resulting alignments were visualized using the genomeVISTA browser, a web –based JAVA applet allowing for the interactive viewing of multispecies sequence alignment for each of the regions submitted. The applet displays a separate track for each species, one below the other which contains a conservation curve based on the selected parameter values. Each of the curves represents the level of conservation observed between the pairwise alignments of each species with the human genome (base genome). In addition, annotation tracks containing known genes, SNPs and repeat regions are provided. Regions under the curve are highlighted using different colours allowing for the identification of conservation of exons, UTRs and non-coding regions.

The alignment of with all 8 species was analysed and regions that exhibited levels of sequence conservation greater than 70% across all species including the zebrafish, fugu and frog were identified. These regions overlapped many annotated genes as was expected, however only non-protein coding regions that displayed high levels of conservation were identified as regions of interest for further analysis. The corresponding human and orthologous sequences, genomic coordinates and alignments for each region were extracted.

2.2.3 In silico analysis of highly conserved regions

The regions that exhibited high levels of conservation across all 8 species were subjected to an *in silico* analysis to determine if they contained any motifs associated with known functional elements. Each region was subjected to a BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) search to confirm the identification of the orthologous sequences and to identify any matches to known human cDNAs. Further characterization of the conserved regions was conducted by submitting the sequences to

various genomic functional annotation tools which aim at predicting whether a sequence contains any functionally relevant signatures associated with known gene and regulatory elements.

2.2.3.1 Prediction of gene structures

Each conserved sequence was submitted to two gene prediction programs, GeneID (http://genome.imim.es/geneid) and GeneScan (http://genes.mit.edu/GENSCAN). Both programs predict genes in genomic sequences using different methodologies. GeneID is designed based on the hierarchical prediction of splice sites and start and stop codons using a Hidden Markov Model. Each site is scored and exons built from these sites, followed by prediction of the final gene structure based on maximizing the sum of scores of the assembled exons. GenScan predicts genes using a similar approach, based on generalized Hidden Markov Model for coding DNA.

2.2.3.2Identification of transcription factor binding sites

Conserved sequences were assessed for transcription factor binding sites by searching against the JASPAR database of known transcription factor binding sites. The JASPAR database contains a collection of all experimentally validated transcription factor binding sites and the factors that bind to them. The database was established as a repository for all transcription factor binding site consensus sequences and position-weighted matrices which are used to estimate the likelihood that a given sequence binds to a specific transcription factor. These matrices have been derived from experimental data on transcription factors and their binding sites. The database was accessed through the ConSite suite of tools. The sequences were first subjected to a single sequence analysis which searches for transcription factor binding sites within a sequence based solely on position weight matrices derived for all previously characterised transcription factors. Each of the sequences were subsequently subjected to a second analysis which searches for major transcription factor binding sites based on the submission of the aligned mouse and human sequences. This approach which is known as phylogenetic footprinting serves to reduce the number of false positive predictions

by only identifying putative sites conserved between the mouse and human sequences. All conserved sequences were assessed for the presence of transcription factor binding sites and the results further analysed to determine the likelihood that the sites found were biologically functionally relevant. This assessment was based on the position of the sites in relation to transcription initiation sequences of genes within the region and the presence of clusters of transcription binding sites know to occur together to bind dimer transcription factors.

2.2.3.3Identification of conserved RNA structures

The prediction of non-coding RNA genes is a difficult process; however there are several algorithms designed to detect regions of pre-mRNA and mRNA predicted to have a conserved secondary structure which may be involved in translational regulation or mRNA localisation or degradation. The prediction of potential micro RNA (miRNA) structure formation of the conserved sequences was carried out using the automated Vienna RNAfold server (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi). Sequences were submitted to the prediction pipeline which predicts potential miRNAs based on the following factors: formation of a stem-loop secondary structure, calculation of free energy, the identification of homologues in relative species and conservation with other known miRNAs. All predicted RNA secondary structures were further assessed to determine their possible biological significance through the analysis of possible gene targets were potential secondary structure formation was identified. Potential RNA genes were subjected to a BLAST search against all known expressed sequence tags and coding sequences to identify possible gene targets and these genes were assessed to determine if they could play a potential role in the etiology of KWE.

2.2.4 Laboratory validation of identified functional regions

In the case where potential functional elements were identified within highly conserved sequences of the KWE critical region, these regions were sequenced by a colleague, Miss Angela Hobbs in affected and unaffected KWE samples to determine if there were any variations that were found exclusively in affected individuals.

2.2.5 Comparative Genomic Hybridisation Array (aCGH)

Comparative genomic hybridization arrays provide a means by which to compare the copy number of differentially labeled test and reference DNA by hybridizing the DNA together to an array containing thousands of probes covering a particular region of the genome. The test and reference samples are labeled with either Cy3 or Cy5, which are fluorescent cyanine dyes with different emission and excitation wavelengths. The measurement of the resulting fluorescence ratios is then plotted probe by probe, in relation to the probe position and provides an efficient means to accurately determine copy number of defined regions. The high resolution capabilities of aCGH allows for the identification of extremely small regions that may show CNV. An aCGH was employed in this study to identify the presence of copy number variants within an extended KWE critical region in KWE affected individuals.

2.2.5.1 DNA Extraction

DNA was extracted from whole blood by diagnostic laboratory personnel using the salting out method. The quality of the extracted DNA was confirmed by running it on a 0.8% agarose gel and the quantification and purity ($OD_{260/280}$) of the samples were ascertained using the Nanodrop ND-1000 Spectrophotometer. The DNA was then aliquoted into final volumes containing 10µg of DNA. The DNA samples were sent to Roche NimbleGen® core facility in Iceland where the aCGH experiments were conducted.

2.2.5.2 Custom CGH tiling array design

A custom tiling array (Roche-NimbleGen) containing roughly 385 000 probes at a resolution of 11bp was designed based on an extended region encompassing the inverted region on chromosome 8p and the KWE-linked critical region between markers D8S1759 and D8S552 (chr8:11 415 050-12 886 829). The array was designed to provide a high resolution tiling path of the entire region and therefore regions known to contain copy number variants and repeat elements were included in the design. The array production and hybridization experiments were conducted by Roche- Nimblegen following their standard protocol. Each of the ten samples were fragmented, labeled and hybridized to a single array together with an equivalent amount of DNA from a differentially labeled reference sample which consisted of a pool of DNA from six Caucasian individuals supplied by Roche-Nimblegen. The use of a pooled reference sample allowed for compensation for the effect of normal copy number variation present in the genome of a single individual affecting the results of the experiment. The same reference sample was used for each of the hybridizations. Following the hybridization experiments, the microarrays were scanned and the fluorescence intensity data extracted.

2.2.5.3 Data Analysis of 385k tiling array

Both normalized and raw datasets were obtained for each experiment from Roche NimbleGen. The fluorescence intensity raw data was normalized for each spot using the qspline algorithm. Normalisation of the data is required in order to compensate for the differences in intensity between the Cy3-labeled test sample and the Cy5-labeled control sample. Qspline normalisation is a quantile-based simple and robust non-linear method for normalisation of two-colour array experiments that has been shown to perform comparably to loess normalisation (Workman et al., 2002). It is the preferred method of normalisation compared to lowess-based methods as it is computationally more efficient, deterministic and can normalise for signal dependant bias across arrays with different features. In addition, a window averaging dataset was obtained using a transformation algorithm to reduce the data

points based on averaging across 10 data points. This dataset therefore contained the average intensities across 100bp (10 data points on average). This dataset provided a means by which to analyse larger regions for CNV. These two datasets were used for further analyses.

Two datasets used:

1. Normalised dataset – 384 702 datapoints
2. 100bp Averaged dataset – 54 632 datapoints

2.2.5.4 Visual Analysis using SignalMap software

Both datasets were visually interrogated using the SignalMap software provided by Roche NimbleGen to identify variation within known genes. The SignalMap software allows for the visualization of the intensity log2ratio for each probe in relation to its position in the genome in the form of signal intensity plots. SignalMap provides a gene annotation track that allows for the examination of CNV is relation to known genes. The criteria for the analysis involved visually inspecting probes covering the length of known genes for variation in copy number based on deviations from the expected intensity for normal copy number. The log2ratio conversion of signal intensities in the case of normal copy number allows for the relative probe intensities to lie close to zero while intensity values representing deleted regions are indicated by negative values and amplified regions are represented by positive values. Regions that exhibited copy number variation overlapping known genes in both affected and unaffected samples were noted.

2.2.5.5. Analysis using TM4 Microarray Software Tool- MultiExperiment Viewer (MeV)

MeV is a versatile multi-experiment tool which is part of a suite of software developed for the analysis of expression microarray data using sophisticated algorithms for clustering, statistical analysis and visualization (http://www.tm4.org/). MeV was further developed to include powerful tools for the visualization and analysis of aCGH data. The tool allows for comparison of samples to identify common regions duplicated or deleted between samples.

2.2.5.6 Analysis of known genes

MeV was utilized to confirm if any CNVs overlapped known RefSeq genes within the tiled region. The algorithm uses an approach which identifies probes which have an intensity log2 value of ≤ -0.8 or ≥ 0.8 for the identification of single copy deletions and duplications respectively and intensity log2 values of ≤ -1 and ≥ 1 for 2 copy deletions and duplications respectively. When searching for significant CNVs within genes, this algorithm is applied across all probes covering the length of a gene.

2.2.5.7 Identification of CNVs present only in affected samples

Since the aim of the experiment was to identify CNVs that may be involved in KWE, regions that are either deleted or duplicated across all affected samples were identified. This was achieved by doing a pairwise assessment of samples based on the comparison of each affected sample with their unaffected relative. The comparison algorithm used a threshold value for CNVs as previously mentioned. In addition, the algorithm searches for regions that exhibit copy number variation (where more than one adjacent probe is assessed for CNV) as opposed to single clones. The algorithm identifies duplicated or deleted regions present in either the affected or unaffected sample or both. CNVs identified in both the affected and unaffected samples were excluded since they are unlikely to be involved in KWE. CNVs identified in only the affected samples were noted and 10 datasets (5 for duplicated regions

and 5 for deleted regions in each affected sample) for the pairwise comparisons were obtained for regions showing either duplications or deletions in each affected sample. These 10 datasets were then compared using the web based tool Galaxy (http://main.g2.bx.psu.edu/) and to identify regions that were commonly deleted or duplicated across the pairwise datasets. This approach provided a means by which to identify CNVs present exclusively in KWE affected individuals and therefore likely candidates to be further assessed for involvement in the etiology of KWE.

2.2.5.8 Assessment of potential functional impact of CNVs

All regions that exhibited copy number variation in both protein coding and non-coding regions were further assessed to determine if they were likely to be involved in the molecular cause of KWE. In the case of CNVs within known or predicted genes, this assessment was based on a literature search to identify whether the established molecular functioning of the genes were associated with the phenotypic manifestations of KWE. Where CNVs were identified within non-coding sequences of the KWE critical region, these were assessed for possible functional impact on the previously identified candidate genes based on their location with respect to these genes. CNVs found upstream or downstream of candidate genes have the potential to be involved in the regulation genes that they lie close to and were therefore further analysed.

Chapter 3. Results

The aim of this study was to examine the KWE critical region on chromosome 8p22-23.1 to identify: (a) highly conserved regions and (b) regions of copy number variation that may be associated with the molecular etiology of the disease. This study first explored the possibility that the pathogenic mutation may lie within a previously uncharacterised highly conserved genomic element in the KWE critical region. Secondly an extended region encompassing the KWE critical region was analysed to identify copy number variation that may be present in KWE affected individuals. Copy number variation has been found to be associated with several genetic disorders and thus may play a role in the molecular cause of KWE. Although both approaches are aimed at analyzing the KWE critical region for possible functional elements related to the molecular etiology of the disease, the two approaches are diverse in their methodology and application and are therefore dealt with separately with interesting regions of overlap being addressed in the discussion.

3.1 Validation of KWE critical region

Although the KWE critical region had previously been identified as lying on the short arm of chromosome 8 the details of the polymorphic inversion that occurs in this region was not clearly defined and therefore the impact of the inversion on the KWE critical region was unknown. Previous KWE studies (Starfield et al., 1997; Appel et al., 2002) together with current studies examining the polymorphic inversion (Giglio et al., 2002; Giorda et al., 2007) revealed that the breakpoints of the inversion had been localized to 2 distinct regions on the short arm of chromosome 8.

Data from the molecular genotyping of microsatellite markers and construction of haplotypes surrounding the KWE critical region generated by a colleague Angela Hobbs, revealed a common haplotype in 4 KWE families. The haplotype analysis was conducted based on the evidence that the KWE critical region lies on an inverted chromosome.

Additional markers not used in previous KWE studies were also incorporated in order to increase the informativity of the haplotypes. By tracking the inheritance of alleles based on Mendelian segregation, the KWE critical region was found to lie in the region between and including markers D8S1759 extending distally towards the REPD and D8Sdi on the inverted chromosome. Lack of additional informative recombination events therefore defined the KWE critical region as lying between markers D8S1819 and D8Sdi on the inverted chromosome. Although a common founder haplotype was present in the affected families studied, the exact boundaries of this haplotype could not be accurately determined. The KWE critical region was previously defined as lying between markers D8S1759 and D8S552 on the uninverted chromosome and was found to encompass the REPP (Figure 3.1) where one of the breakpoints of the inversion is known to occur. In the presence of the inversion the critical region will still appear to lie between the two previously mentioned markers, however the region is significantly increased to approximately 6Mb encompassing many more genes (Figure 3.1).

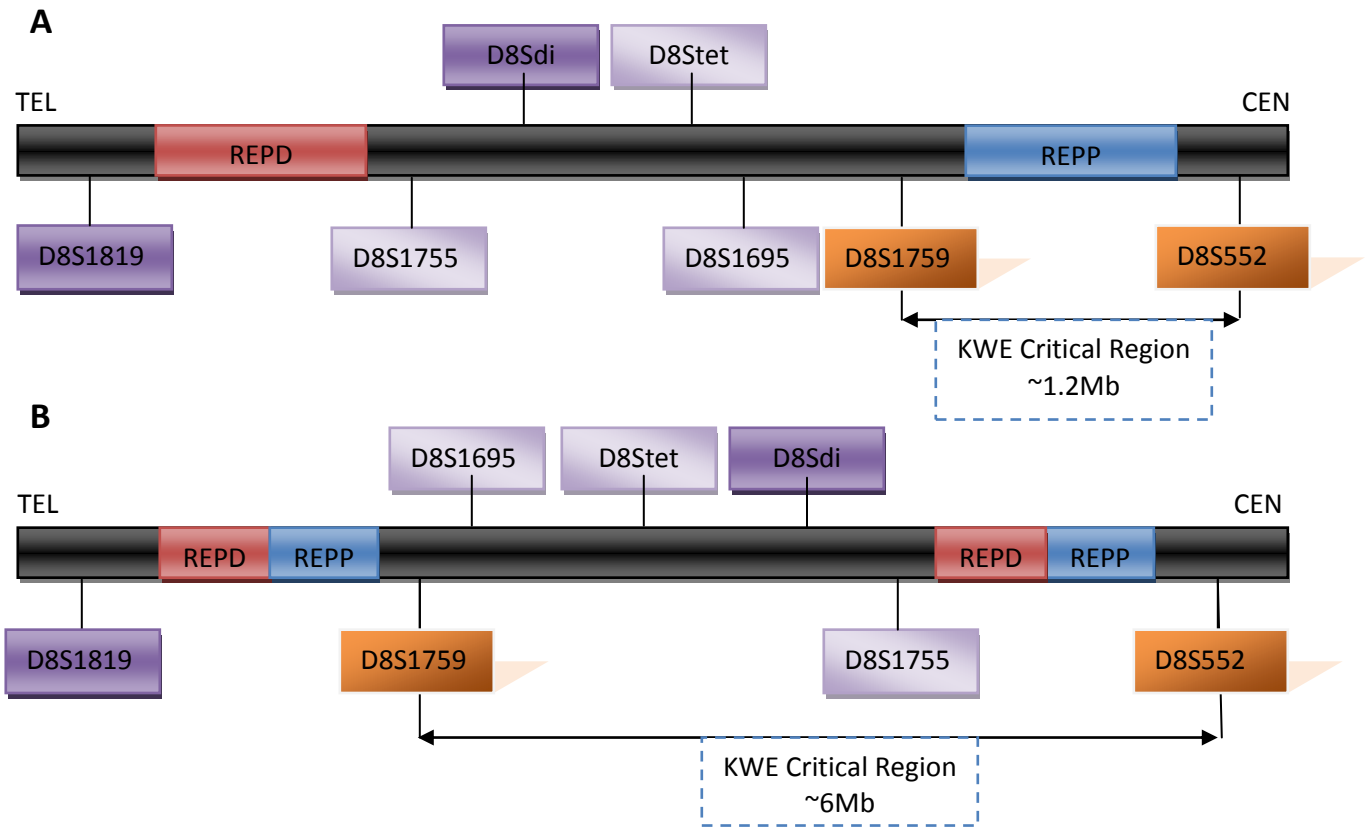


Figure 3.1 Schematic representation of the position of the KWE critical region. (A) Shows part of chromosome 8p in the normal orientation, whilst (B) shows the altered position of the KWE critical region on the inverted chromosome based on the haplotype analysis.

Although the additional haplotype analysis revealed a possible new KWE critical region on the inverted chromosome, the markers that were typed could not provide accurate information regarding the boundaries of this new region. Further markers would need to be typed in future studies between markers D8Sdi and D8Stet and D8S1819 and D8S1759 in order to define the boundaries. Due to the lack of more informative information from the haplotype analysis, the results from previous studies were used to assess the region to be analysed for conservation. In the absence of additional information to define this new critical region as observed on the inverted chromosome, the original KWE critical region between markers D8S1759 and D8S552 was used as the starting point to identify highly conserved regions. The region analysed for conservation was therefore defined as lying on the uninverted chromosome 8 approximately 200kb upstream of marker D8S1759 and 100kb downstream of D8S552 at position 11,317,148-12,899,353 (Mar.2006, HG18). The flanking regions were included so as to ensure that the sequence surrounding the markers was also included in the analysis.

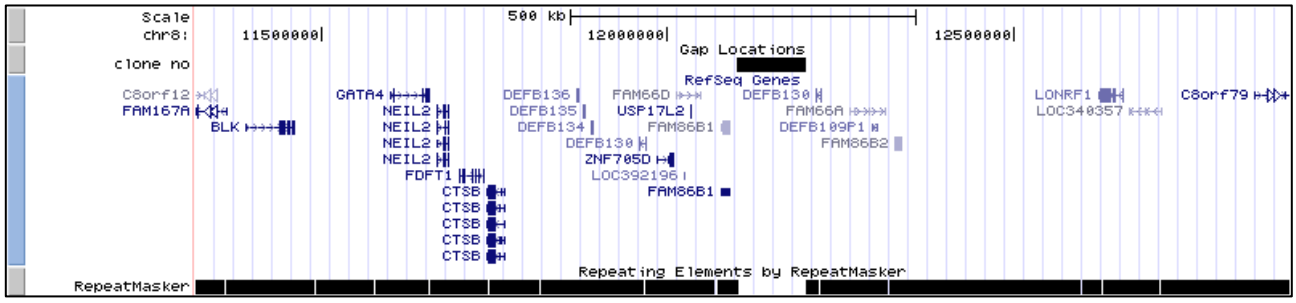


Figure 3.2 A representation of the KWE critical region between markers D8S1759 and D8S552 showing RefSeq annotated genes in the region (UCSC Genome Browser – Accessed January 2010).

Figure 3.2 shows the annotated genes within the region between markers D8S1759 and D8S552. The region is known to contain 8 protein coding genes, a beta-defensin gene cluster and a number of additional hypothetical genes and open reading frames. In addition, the region also contains a clone gap in the assembly of the human genome sequence which is approximately 100kb in size.

3.2 Comparative genomic analysis of KWE critical region

A number of analytical approaches have been developed to address the challenge of identifying conserved regions across multiple orthologous sequences. Each of these approaches has attempted to take into account a number of confounding factors which can dramatically affect the outcome of the alignment of sequences from multiple species. The KWE critical region was aligned with 8 evolutionarily diverse organisms in order to identify regions that were highly conserved. The 8 organisms were selected on the basis of representing both closely and distantly related organisms; hence the regions of interest were those that exhibited a high level of sequence conservation across a majority of the most evolutionarily diverse species. The comparison of DNA sequences from multiple species that range in evolutionary distances allows for the identification of not only slowly evolving coding sequences but also the conserved non-coding elements that may harbor important functional or regulatory elements. For this reason, the species used in the comparative genomic analysis ranged from the human and pufferfish which diverged approximately 450 million years ago to the human and mouse which diverged from a common ancestor approximately 40 – 80 million years ago. A significant number of conserved regions were observed, with majority of the conservation occurring within the exons of annotated protein coding genes in the region. Although these regions were highly conserved, this was expected and served to prove the premise that functional regions in the form of protein coding genes are under selection to remain unchanged, even between distantly related organisms.

3.2.1 Identification of highly conserved regions

To optimize the chance of identifying all possible non-coding conserved regions the GenomeVISTA server uses a computational strategy whereby the query sequence contigs are anchored to a base genome using a local approach followed by a global alignment of larger anchor regions. In order to identify regions within the KWE critical region that were of potential functional relevance, genomic regions which were at least 100bp in length and conserved at a level of greater than 70% across species were identified. Based on these

criteria a large proportion of the KWE critical region was found to be conserved between the closely related organisms such as the primates and mammals. This was expected as the majority of evolutionarily conserved regions are under constraint to remain unchanged and thereby preserve their functionality. The conservation of the previously studied KWE candidate genes in the region were of significant interest, since both *FDFT1* and *CTSB* showed high levels of conservation across all 8 of the species (Figure3.3). Although the main focus of this study was to identify non coding regions, additional regions of interest in potential coding regions were also investigated due to their high levels of conservation between species.

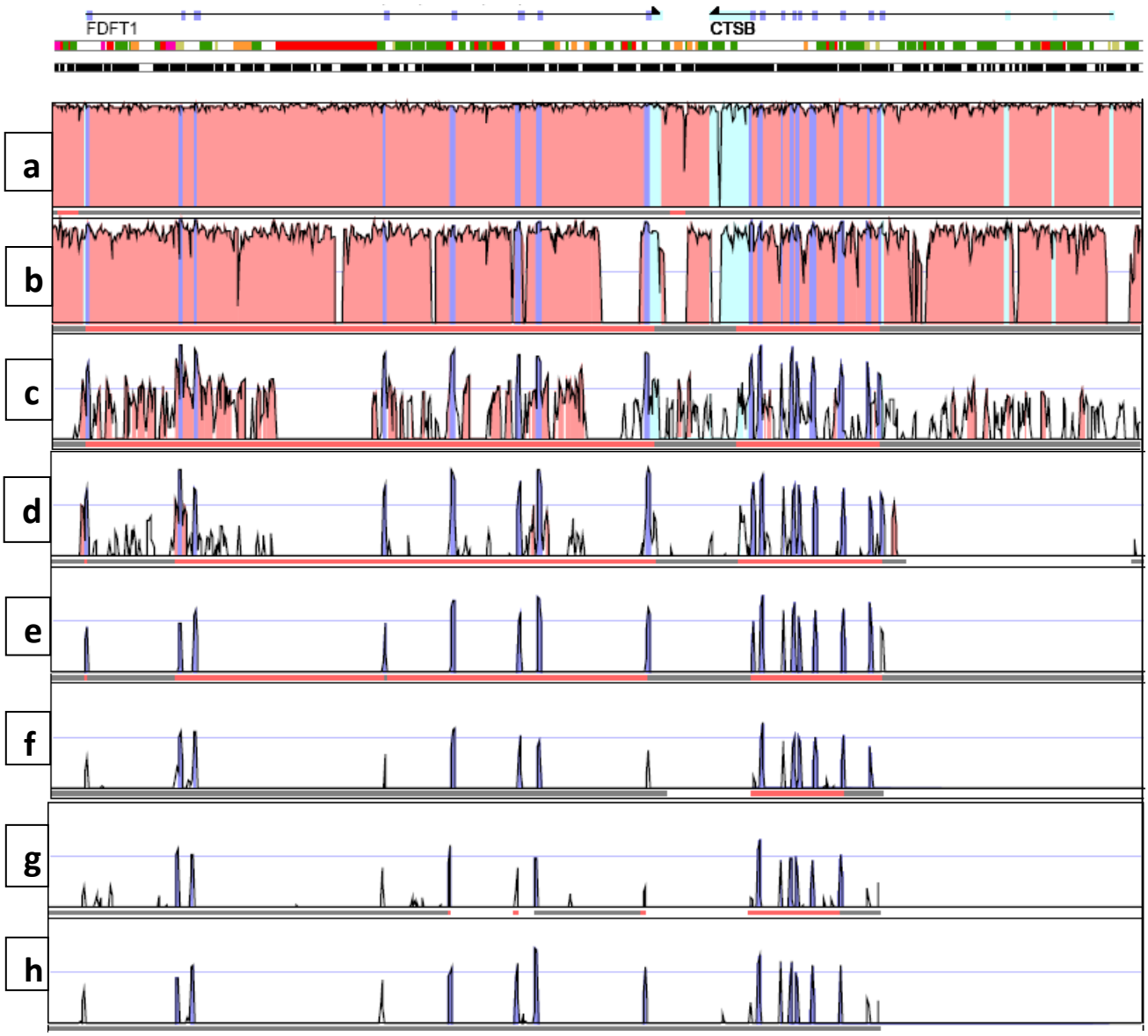


Figure 3.3 GenomeVISTA conservation plot showing the two KWE candidate genes *FDFT1* and *CTSB*. The blue peaks represent exons that show conservation of greater than 70%, while the pink regions represent non-coding conserved regions. Each track is the alignment of the human genome sequence with (a) Chimp, (b) Rhesus, (c) Dog, (d) Mouse, (e) Chicken, (f) Fugu, (g) Zebrafish and (h) Frog.

Eight distinct regions of significance were identified in the multiple sequence alignment. The identification of these regions was based on the analysis of the GenomeVISTA multiple sequence alignments. The level of conservation between the human genomic sequence and all other sequences were assessed across the entire KWE critical region in 300kb segments. Regions exhibiting levels of conservation above the 70% threshold were indicated by peaks in the GenomeVISTA display (Figure 3.3). Of these eight regions, six were identified within non-

coding regions of the genome while the remaining two were identified within a previously annotated open reading frame. Each of these regions ranged in size from 100bp to 406bp (Table 3.1). These regions were selected as regions of interest based solely on the fact that they shared orthologous regions in at least 6 of the most evolutionarily diverse species which were included in the multispecies alignment.

3.2.2 Assessment of highly conserved regions

Regions 1 and 2 were found to be conserved across all 8 species within a region that had been characterised as a predicted open reading frame *c8orf13* (Figure 3.4). As previously mentioned, although this region was identified within a possible coding region of the genome, the high level of conservation presented it as a region of interest for further analysis. Region 3 was a large intergenic non-coding region of 406bp which was conserved in 6 of the 8 species inclusive of the Zebrafish and *Fugu*. This region was situated between two known RefSeq genes *BLK* and *GATA4* and therefore could possibly be involved in the regulation of these genes. Regions 4 and 5 were two smaller conserved regions within approximately 4500bp of each other. These intergenic non-coding regions were 154bp and 124bp in size respectively and were found to be conserved in 6 species inclusive of the evolutionarily distant *Fugu* and Zebrafish (Figure3.5). Similarly regions 6 and 7 were two small regions found within 537bp of each other which were conserved in 6 of the species in the alignment (Figure 3.5). At the time that the alignment was first conducted, these two regions were identified as non-coding intergenic regions although in figure 3.5 the region is shaded in blue indicating a protein coding exon. The reason for this will be addressed later in section 3.2.3.1. The close proximity of regions 4 and 5 and regions 6 and 7 to each other respectively presented them as interesting potential functional candidates. Region 8 was found to be conserved in 6 of the 8 species including the *Fugu* and Zebrafish and spanned 262bp. Interestingly, this region showed high levels of conservation across the two primates and dog, but was not conserved in the mouse and the chicken.

The length of the conserved sequence for each of the regions varied between organisms. The lengths indicated in table 3.1 are the longest regions of conservation that was observed for

each of the regions. The corresponding coordinates and genome builds for each of the conserved regions in the different species analysed in the multiple alignment is shown in appendix C.

Table 3.1 Human chromosomal coordinates of regions showing high levels of conservation across at least 6 species with a cutoff of 70% over 100bp

	Region	Length	%Conservation
1	Chr8: 11 319 292 – 11 319 555	264bp	70 – 98
2	Chr8: 11 338 950 – 11 339 330	381bp	70 – 99
3	Chr8: 11 512 454 – 11 512 859	406bp	76 – 98
4	Chr8: 12 023 858 – 12 024 011	154bp	71 – 90
5	Chr8: 12 028 566 – 12 028 689	124bp	71 – 92
6	Chr8: 12 032 672 – 12 032 771	100bp	70 – 98
7	Chr8: 12 033 308 – 12 033 490	183bp	72 - 95
8	Chr8: 12 412 983 – 12 413 244	262bp	70 – 94



Figure 3.4 Genome VISTA conservation plot for the open reading frame c8orf13. The red boxes clearly show the conservation of the exons in the hypothetical open reading frame c8orf13. Each track is the alignment of the human genome sequence with (a) Chimp, (b) Rhesus, (c) Dog, (d) Mouse, (e) Chicken, (f) Fugu, (g) Zebrafish and (h) Frog

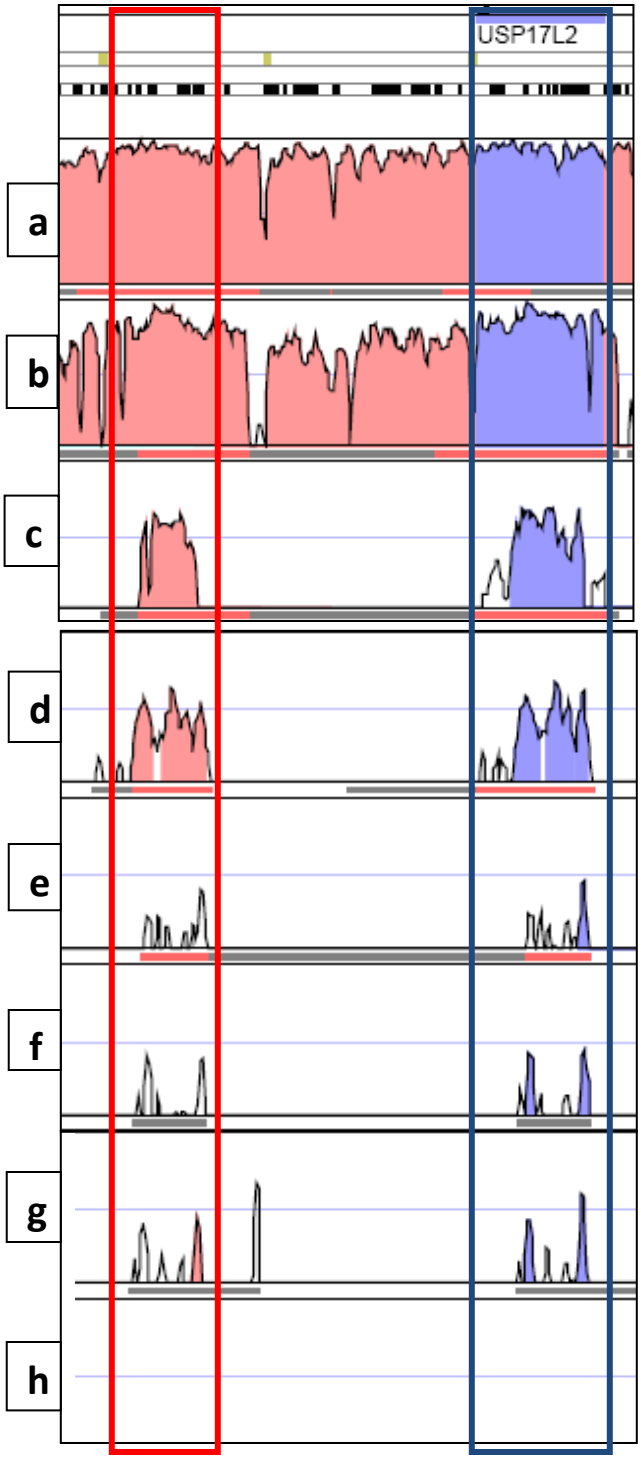


Figure 3.5 GenomeVISTA plot showing conservation of non-coding regions 4 &5 (red box) upstream of regions 6&7 (blue box) that showed two distinct regions of sequence conservation within a hypothetical gene which was previously uncharacterised. Each track is the alignment of the human genome sequence with (a) Chimp, (b) Rhesus, (c) Dog, (d) Mouse, (e) Chicken, (f) Fugu, (g) Zebrafish and (h) Frog

3.2.3 Functional analysis of highly conserved regions

The evolutionary conservation of the regions identified above presented them as potential functional elements in the KWE critical region. Therefore each of the eight highly conserved regions was assessed for potential functional roles via three *in silico* approaches; gene structure prediction, RNA secondary structure formation and the prediction of transcription factor binding sites. There are several computational approaches which have been developed to address the challenges associated with predicting *in silico* functional elements in the genome. These methods use the information gathered from experimentally validated functional elements to create models and algorithms that allow for the prediction of functional elements. Although these approaches have matured significantly in recent years resulting in the identification of several experimentally validated functional elements in the genome, the prediction of vast amounts of false positives remains a significant challenge. In order to address this problem the prediction approaches and tools used in this study were those found to be most robust to minimise the occurrence of false positive results.

3.2.3.1 Prediction of gene structures

The identification of possible gene structures in the conserved sequences was conducted using two of the most accurate gene prediction programs, GENSCAN and geneID. Since the identified conserved regions may have formed part of exons within possible gene structures, 1000bp of flanking sequence upstream and downstream was included for the analysis of each region. As previously mentioned, the close proximity of conserved regions 4 & 5 and 6 & 7 presented them as possible exons within a larger gene structure. Similarly, regions 1 and 2 had previously been identified as c8orf13 and therefore these regions were analysed independently as well as together as one large sequence. Both GENESCAN and geneID aim to create prediction models that capture the general and specific compositional properties of the functional units of a eukaryotic gene which include exons, splice sites and promoters. Table 3.2 shows the results obtained from the analysis of the regions using both programs.

Table 3.2 Gene structure prediction results showing the highest scoring gene predictions for each of the conserved regions using the two different prediction tools, GENESCAN and GeneID

	GENSCAN Prediction					GeneID Prediction			
Region	**Genomic Coordinates**	**No. of exons**	**Length of coding sequence (bp)**	**Length of AA sequence (aa)**	**Exon Type**	**No.of exons**	**Length of coding sequence(bp)**	**Length of AA sequence(aa)**	**Exon Type**
1	chr8:11318292 - 11320555	1	264	87	Terminal	1	264	88	Terminal
2	chr8:11337950 - 11340330	3	684	228	Internal	1	381	127	Initial
1 & 2	chr8:11318292 - 11340330	3	645	214	Initial Terminal, PlyA	2	282	94	Initial Terminal
3	chr8:11511454-11513859	1	198	66	Internal	1	159	53	Internal
4	chr8:12022858-12025011	2	492	164	Initial Internal	2	525	175	Initial Internal
5	chr8:12027566-12029689	1	1248	416	Initial	1	1248	416	Initial
6	chr8:12031672-12033771	1	1593	530	Single	1	1593	531	Single
7	chr8:12032308-12034490	1	1371	457	Initial	2	1305	435	Initial Internal
4,5,6,7	chr8:12022858-12034490	2	1593	530	Single, PlyA	1	1593	531	Single

8	chr8:12411983-12414244	1	120	40	Internal	None

Despite advances in the field of computational gene prediction, the methods employed in gene-finding are far from being able to accurately predict the exonic structure of genomic sequences due to the prediction of a large number of false positive exons and genes. The combination of two of the most widely used methods was therefore adopted in order to compare the results obtained to address this issue.

3.2.3.1.1 Gene structure in regions 1-3

Regions 1 and 2 were previously shown to be conserved sequence within a characterised open reading frame *c8orf13.* The function of this open reading frame is yet unknown; however it has been shown to be expressed and was previously characterised as a hypothetical protein. Based on the gene structure prediction, region 1 showed evidence for a possible single terminal exon (3' splice site to stop codon), while region 2 showed a possible gene structure consisting of 3 internal (3' splice site to 5' splice site) exons. The prediction for regions 1 and 2 together with the intervening sequence produced varying results for each of the programs. GENESCAN predicted 3 possible exons consisting of an intial (ATG to 5' splice site) exon, terminal exon and polyA tail, while GeneID predicted an initial and terminal exon. In this case, the position number and lengths of the exons varied significantly between the two programs. The likely reason for this is based on the fact that *c8orf13* has previously been characterised as containing 3 exons with the region analysed only encompassing the last 2 exons and partial fragment of the first exon. In addition, the region is also known to overlap with another characterised open reading frame *c8orf12*, although the conserved regions fall within the intronic sequence of this open reading frame. Gene-finding programs such as GeneID have been shown not to be able to deal with uncommon phenomena such as nested genes, alternate splicing and genes within introns.

Interestingly at the time that the study was conducted, the region had been annotated as open reading frame *c8orf13*. Subsequently the region has been validated as a hypothetical protein coding gene and renamed *Homo sapiens family with sequence similarity 167, member A (FAM167A).* Although this region initially presented as an interesting positional

candidate since little is known about its function, it was located significantly upstream of the boundary marker D8S1759 for the KWE critical region and had been examined for variations in a previous KWE study (Appel et al., 2002). None was detected to co-segregate with KWE. The gene structure prediction for region 3 revealed a single internal exon within the region using both programs. The score for the prediction was extremely low in both instances and since the predicted exon was that of an internal exon with no initial or termination signals, the prediction was unlikely to be real, although it could possibly be part of a larger gene structure. This was however unlikely based on the fact that the surrounding regions did not show any evidence for sequence conservation.

3.2.3.1.2 Gene structure in regions 4 – 8

As previously mentioned, the close proximity of these conserved regions to each other presented them as likely candidates for possible exons within an extended gene structure. Individually regions 4 and 5 were both predicted to contain a gene structure consisting of 2 and 1 exons respectively. In the case of region 4, both an initial and internal exon was predicted of similar size encompassing the same region using both programs. The gene structure did however lack any signal for a terminator exon and therefore scored poorly for possible overall gene structure in both approaches. In this instance however it could be likely that the region formed a partial segment of a gene and the terminator signal may be found further downstream. Similarly region 5 revealed an identical prediction of a large initial exon of 1248bp using both prediction approaches.

Region 6 presented the most interesting result, with both programs predicting a possibly single-exon gene structure. This prediction is based on the identification of an ATG start codon followed by a significantly long exon terminating in a stop codon. The predicted exon was found to be 1593bp in length, encoding a peptide product of 531 amino acids. Region 7 was predicted to contain 1 and 2 exons respectively based on each programs prediction, although both predictions produced similar coding sequence lengths and predicted peptide sequences. Again here it must be noted that due to the close proximity of regions 6 and 7 to

each other, the addition of the flanking 1000bp resulted in an overlap of sequence between the two regions. Due to this factor, the entire region encompassing regions 4 – 7 was also analysed for gene structure using both programs. Interestingly the highest scoring significant result obtained for the larger region using both approaches was for the single-exon gene structure predicted previously for region 6. In addition, GENESCAN predicted a polyA tail in addition to the single-exon structure. This presented region 6 as a strong candidate for a possibly novel gene within the KWE critical region.

Analysis of region 8 using GENESCAN revealed a low scoring internal exon of only 120bp encoding a predicted peptide of only 40 amino acids, whilst GeneID failed to predict any gene structure for the region. The GENESCAN prediction was therefore likely to be a false positive result and the region is therefore unlikely to contain any elements associated with a functional gene.

3.2.3.1.3 Further analysis of region 6-7

Region 6 -7 therefore presented as a possible functional candidate based on the presence of a predicted single-exon gene structure. Although single-exon genes are more abundant in prokaryotes, a significant percentage has been identified in multi-cellular organisms including humans. In order to determine if there was experimental evidence for the transcription of region 6, the sequence was subjected to a BLAST search against the NCBI Transcript Reference Sequence database (http://blast.ncbi.nlm.nih.gov/Blast.cgi). An exact hit was identified against an mRNA sequence described as *Homo sapiens deubiquitinating enzyme 3 (DUB3)*. The alignment of the region to the mRNA was an exact match represented by 100% identity over the length of the query sequence. There was an additional slightly lower scoring hit with a maximum identity of 96% to an mRNA sequence described as *Homo sapiens ubiquitin specific peptidase 17 (USP17)*. Interestingly *USP17* shared a high sequence similarity to region 6-7 with exactly the same mRNA sequence length with only 4% differing between the two sequences. *USP17* had been mapped to a region on the short arm of chromosome 4 and characterised as a functional deubiquitinating enzyme gene. The high similarity of *DUB3* suggested that the presence of the significantly similar sequence on

chromosome 8 might have been due to a rearrangement or transposition of *USP17* within the human genome. However since region 6 exhibited an exact match to *DUB3* mRNA coupled with a slight variation in sequence similarity to *USP17* this provided sufficient evidence to suggest that the region on chromosome 8 was actively transcribed to produce an mRNA.

DUB3 was first cloned by Burrows *et al.*(Burrows et al., 2004) and identified as a cytokine inducible deubiquitinating enzyme. The gene had however not been mapped to the human genome prior to this study due to the presence of a number of highly homologous sequences throughout the human genome. The prediction of the single-exon gene based on the conserved genomic sequence of region 6-7 coupled with the match to the experimentally validated mRNA sequence provided significant evidence to suggest that the mRNA sequence for *DUB3* mapped to this region. Region 6-7 therefore presented as possible novel protein coding region within the KWE critical region.

Deubiquitinating enzymes or DUBs function by removing the covalently attached ubiquitin from proteins, thereby controlling substrate activity and/or abundance. The mechanisms and targets of most DUBs as well as their regulation is still poorly understood, however it was shown that *DUB3* is expressed as mRNA in many different tissues including the heart, spleen, kidney, colon, skeletal muscle and liver. *DUB3* is a cytokine inducible protein expressed transiently and is thought to play a role in blocking cell proliferation and inducing apoptosis through the ubiquitin pathway. The molecular function of *DUB3* therefore could be involved in the observed KWE phenotype and this presented it as a likely novel candidate gene for KWE.

Subsequently during the duration of the study *DUB3* was annotated to region 6-7 within the KWE critical region on the human genome and was found to consist of a single exon lacking a 5’ or 3’ UTR and encoding a protein of 530 amino acids. This result correlated perfectly with the gene prediction obtained previously for region 6. The newly defined *DUB3* was situated within the defined KWE critical region however it was previously not examined in other KWE studies. Further validation of this region revealed that due to the fact that it shared a

significant amount of sequence similarity with *USP17* and had likely occurred due to a rearrangement or transposition of *USP17*, it was subsequently renamed Homo sapiens ubiquitin specific peptidase 17-like 2 (*USP17L2).*

3.2.3.2 Prediction of transcription factor binding sites

Since regions 1 & 2 and 6 & 7 were found to represent *c8orf13* and *DUB3* respectively they were excluded from the analysis for the prediction of transcription factor binding sites (TFBS). Each of the remaining conserved regions was subjected to a computational analysis to identify possible regulatory regions represented by transcription factor binding sites. As previously mentioned, although there are a number of tools available for the prediction of possible transcription factor binding sites within a DNA sequence, these tools are hindered by the prediction of a vast amount of false positives. To address this problem, a combination of two prediction methods was used to identify possible transcription factor binding sites within the remaining 4 conserved regions. Predictions were first conducted for each region using ConSite which predicts possible TFBS based on the JASPAR CORE database which contains curated, non-redundant matrix models for experimentally validated TFBS in humans. This prediction is based purely on the use of matrix models describing the binding preferences of transcription factors based on known sites for a particular DNA binding protein. The model matrices for all known human transcription factors were used to scan each region to predict TFBS. The top 5 highest scoring TFBS were noted from these results for each region (Table 3.3). The second approach identified TFBS using the JASPAR CORE database for humans TFBS, however this prediction was based on a phylogenetic footprinting approach whereby the human and mouse sequence were used to identify TFBS within conserved regions.

Table 3.3 Transcription factor binding site predictions for conserved regions

	ConSite Single Sequence Prediction				ConSite Phylogenetic Footprinting Prediction			
Region	**Transcription Factor**	**Sequence Motif**	**Score**	**Strand**	**Transcription Factor**	**Sequence Motif**	**Score**	**Strand**
Region 3	c-REL	GGAAATACCA	10.391	-	SOX9	CCATTGTCC	10.380	-
	P65	GGAAATACCA	10.192	-	SRY	ATTGTCCAC	7.585	-
	SOX-9	CCATTGTCC	10.380	-	GATA-3	TTATCA	6.894	-
	Thing1-E47	CTGCCAGATC	10.551	-	GATA-2	TATCA	5.778	-
	SP1	ACCAAGCCTC	9.172	-				
Region 4	SP1	GGGGCGGTGT	10.535	+	SP1	GGGGCGGTGT	10.535	+
	Hen-1	CCGCAGCAGGTC	9.170	+				
	c-REL	GGTATTTCCC	8.954	-				
	p-65	GGTATTTCCC	8.783	-				
	SPI-1	GGGAAG	8.579	+				
Region 5	GATA-3	CTATCT	8.388	-	SP1	GCGGCAGTGT	8.274	+
	SP1	GCGGCAGTGT	8.274	+				
	p65	AGGTATTTCC	8.255	+				
	c-REL	AGGTATTTCC	8.104	+				
	NRF-2	AGCGGCAGTG	8.058	+				
Region 8	Myf	CAGCAGCAGCTG	15.399	+	NRF-2	AGCGGAAGCC	10.103	+
	Myf	CAGCAGCAGCTG	14.747	+	SPI-B	AGCGGAA	9.933	+
	Tal1beta-E47S	GCCAGATGGTCC	12.862	-	SAP-1	AGCGGAAGC	9.282	+
	Myf	CAGCAGCAGTTG	11.939	+				
	Myf	CAGCAGCAGTTG	11.963	-				

The prediction of possible TFBS using the JASPAR CORE database produced numerous putative TFBS for each of the regions. This was expected since this type of analysis has a high sensitivity but a complete lack of selectivity since the prediction is based purely on the identification of sites in a single genomic sequence based on the matrix models for each transcription factor. Although the sites are defined binding motifs for known transcription factors they may not have any biological function *in vivo*. The phylogenetic footprinting approach adds a secondary layer of prediction based on the identification of TFBS occurring only in highly conserved orthologous regions.

In the case of all the regions analysed it was found that the footprinting method significantly reduced the number of putative sites predicted. Region 3 was predicted to contain a number of different sites with the highest scoring prediction for the single sequence analysis revealing a predicted c-REL and SOX9 TFBS. The footprinting approach predicted the highest scoring site to be a SOX9 binding site together with possible SRY, GATA-3 and GATA-2 sites. The footprinting approach failed to predict a number of the higher scoring sites predicted by the single sequence analysis. Region 3 occurs in a region approximately 90kb upstream of *GATA-4* and approximately 54kb downstream from *BLK*. Although the region is quite a distance from these two genes these TFBS could possibly play a role in their regulation.

In the case of regions 4 and 5 only a single TFBS, SP1 was predicted to occur in the region using the footprinting approach. For region 4 this had been the highest scoring site using the single sequence approach; however for region 5 the highest scoring site, GATA-3, failed to be predicted using the footprinting approach. Regions 4 & 5 occur approximately 3.5kb upstream from the *DUB3* and approximately 19kb downstream from *ZNF705D.* SP1 is a common TFBS found within several documented promoter regions and the presence of this binding site could possibly play a role in the regulation of either of these two genes.

Region 8 showed conflicting predictions using the 2 approaches. The single sequence analysis revealed an overrepresentation of a number of Myf TFBS throughout the region together with a Tal1beta-E47S site. In contrast the footprinting approach failed to identify any Myf TFBS but instead predicted NRF-2, SPI-B and SAP-1 binding sites. Region 8 is located in a gene poor region of the genome with the closest functional element being a predicted gene *FAM86B2* approximately 75kb upstream of the region. Together with the conflicting binding site predictions it appears unlikely that the region contains any biologically relevant TFBS.

3.2.3.3 RNA secondary structure prediction

One of the main criteria for the identification of novel miRNA structures is evidence for the expression of a particular genomic sequence. Of the 8 highly conserved regions identified, the only two regions which showed experimental evidence for the expression of a validated mRNA sequence were regions 1 & 2 and regions 6 & 7. Region 1 & 2 were found to encompass part of a characterised ORF, *c8orf13*, which has subsequently been classified as a hypothetical protein-coding gene *FAM167A,* while region 6 & 7 has been identified as a known protein coding gene *DUB3.* Therefore it was unlikely that any of these four regions were likely to be involved in the formation of secondary RNA structures and miRNAs. Of the remaining conserved regions, region 3 showed no experimental evidence for expression, while a number of cDNAs were found to map to the region encompassed by regions 4 and 5 and region 8. These cDNAs were however significantly longer in length, with each of the conserved regions covering a very small portion of the cDNA. These conserved regions were therefore analysed to assess the possibility of secondary RNA structure formation.

3.2.3.3.1 RNA secondary structure in region 4, 5 and 8

The RNAfold server from the Vienna RNA web suite of tools was used to predict the secondary structure for each of the conserved regions. In each case the partition function, base pairing probability matrix and minimum free energy (MFE) structure were predicted based solely on the RNA sequence. One of the defining criteria to be met for the identification of functionally relevant miRNAs is the identification of a predicted minimum free energy fold-back precursor with extensive base-pairing in the miRNA region which should not contain any large internal loops or asymmetric bulges. The lower the free energy of a structure, the more likely it is that the structure is formed, since energy is released. Determining the base-pair configuration with the minimum possible free energy forms the basis of secondary structure prediction algorithms. Like most *in silico* approaches it is prone to an overrepresentation of false positives; however there are a number of additional

algorithms that have been developed to increase the confidence in secondary structure predictions.

Region 4 was predicted to have a MFE of -50.80 kcal/mol producing a secondary structure represented in figure 3.6. The corresponding “positional entropy” for the structure is indicated by the varying colours in the thermodynamically stable MFE structure ranging from red (low entropy, well-defined) via green to blue and violet (high entropy, poorly defined). It is clear that there exists a strongly predicted possible stem-loop region on the right of the secondary structure, however the rest of the secondary structure appears to be poorly predicted based on the positional entropy. Although the stem-loop structure is confidently predicted it lacks the required amount of nucleotides within the stem to possibly be processed into a functional miRNA.

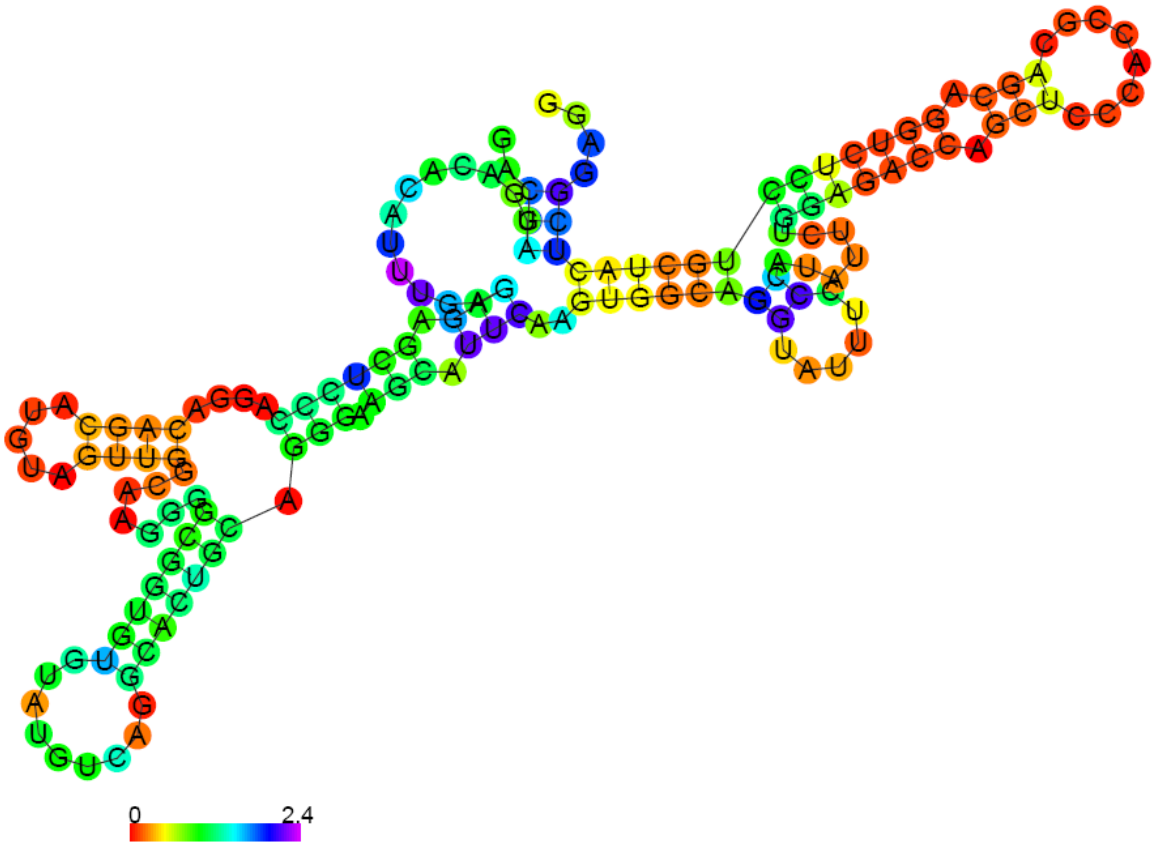


Figure 3.6 Minimum free energy predicted secondary structure for region 4

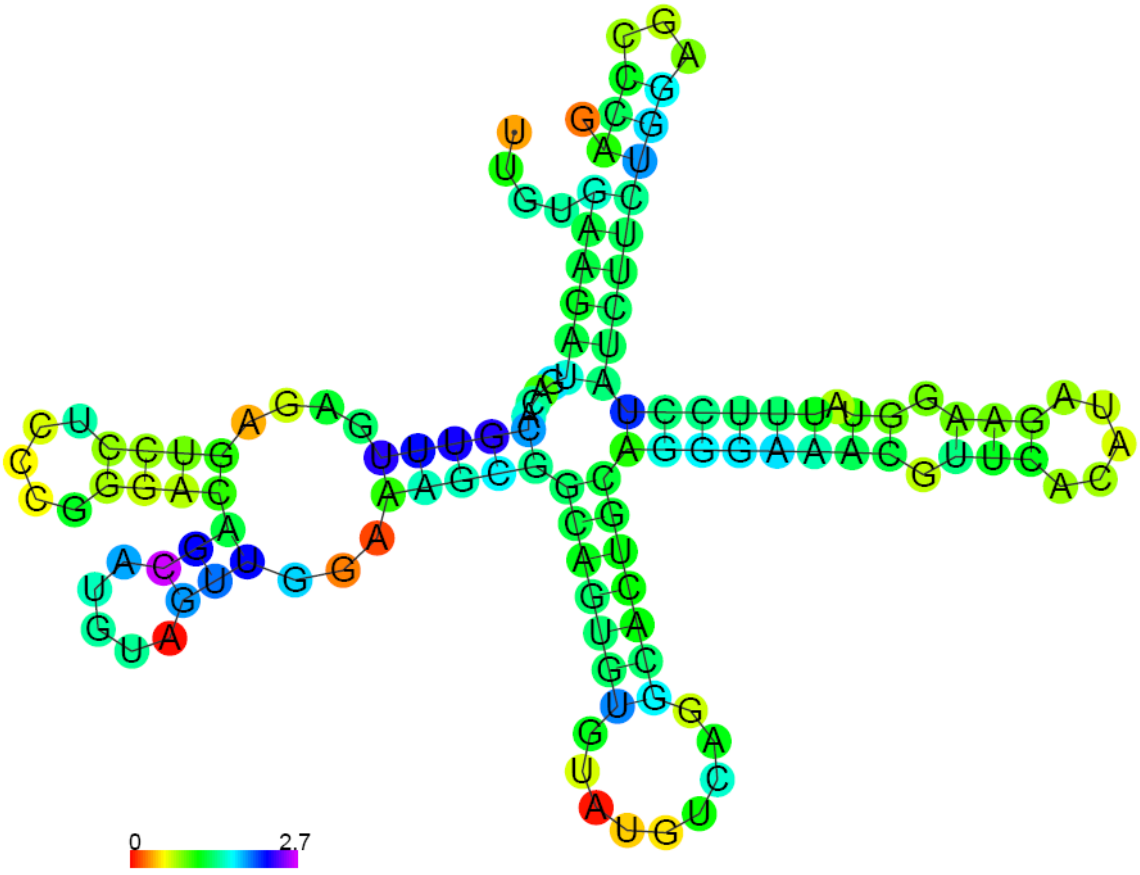


Figure 3.7 Minimum free energy predicted secondary structure for region 5

Region 5 was predicted to produce a secondary structure with a MFE of -37.43 kcal/mol. A number of stem and loop structures were predicted within the secondary structure; however the positional entropy of these regions failed to support the MFE structure with most of the structure exhibiting high entropy. The prediction of the secondary structure was therefore most likely to not be biologically relevant.

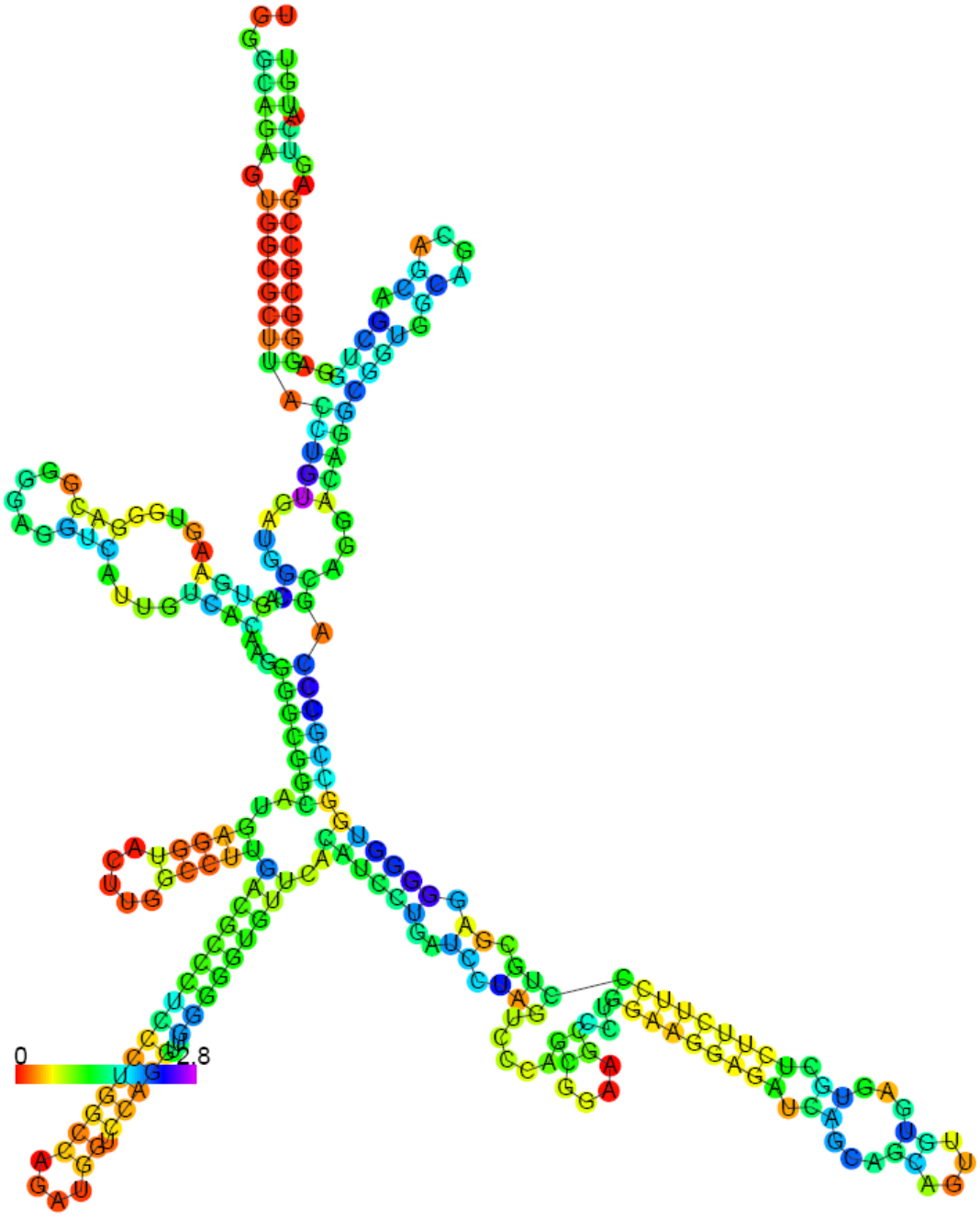


Figure 3.8 Minimum free energy predicted secondary structure for region 8

Region 8 displayed a prediction of a number of stem-loop structures in an overall MFE secondary structure of -119.23 kcal/mol. The overall secondary structure exhibited a very low MFE however the stem-loop regions varied significantly in the level of positional entropies providing little confidence in the prediction.

3.3 Comparative genomic hybridization array of KWE critical region

The second part of this study aimed to identify regions of copy number variation within an extended KWE critical region. It is now known that copy number variation of specific regions of the genome contributes to the neutral variation observed in all individuals. It therefore remains a challenging task to identify CNVs that may contribute to a disease phenotype.

3.3.1 Copy number variation within genes

Visual analysis of copy number variation across the tiled region revealed several regions that exhibited deletions or duplications (Table 3.4) in individual samples. Closer analysis of these regions showed that most variation was observed within the olfactory repeat region which also harbours the gene family of the beta-defensins in a repeat cluster. As previously mentioned two olfactory repeat clusters are found within the tiled region and flank the two sequence gaps, a proximal (REPP) and distal (REPD) repeat region (Figure 3.9). These two repeat clusters are implicated in a large inversion (3.5Mb) that occurs within the region which has been observed in KWE affected individuals. This inversion is polymorphic (Giglio et al., 2001; Shimokawa et al., 2004). In addition to the beta-defensin genes, there are a number of other hypothetical genes that also exhibited copy number variation including *FAM90A*, *SPAG11A* and *FAM90A25P*. *FAM90A* is a primate-specific gene family having thought to have arisen due to multiple duplications and rearrangement events, while *SPAG11A* is a sperm associated antigen. The beta-defensin gene cluster has previously been shown to exhibit variations in copy number (Figure 3.9).

Although the regions mentioned in Table 3.4 did show evidence for varying copy number, no duplications or deletions were identified that were consistent in all five KWE affected individuals. In addition, these regions all fell significantly outside of the KWE critical region. All well characterized RefSeq genes, including those genes within the KWE critical region, did not show any CNVs based on the visual analysis as well as the analysis using the MeV software based on the previously defined thresholds.

Table 3.4 Regions identified as displaying copy number variation by visual analysis of normalized and 100bp averaged probe intensity plots in all 10 samples using SignalMap. These regions exhibited variation from the normal copy number and were either duplicated or deleted

Genes	Location
Defensin A1\Defensin A3	6 822 602 – 6 863 213
FAM90A-Pseudogene cluster	7 101 853 – 7 140 853
DefensinB 106A	7 327 453 – 7 331 319
DefensinB 107A	7 340 777 – 7 354 343
FAM90A-Pseudogene cluster	7 393 430 – 7 427 585
DefensinB 106A	7 720 103 – 7 723 935
SPAG11A	7 742 811 – 7 758 729
FAM90A25P	12 316 403- 12 322 770

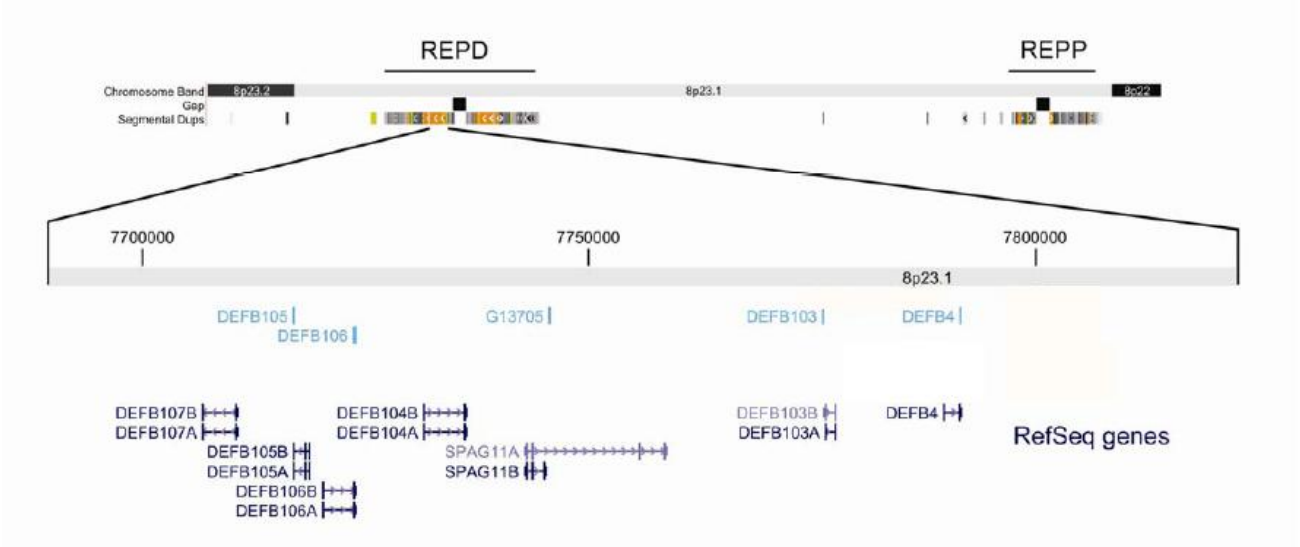


Figure 3.9 Part of tiling array showing the olfactory repeat regions REPD and REPP, gap positions and segmental duplications highlighted. The beta-defensin gene cluster within the olfactory repeat region is expanded below.

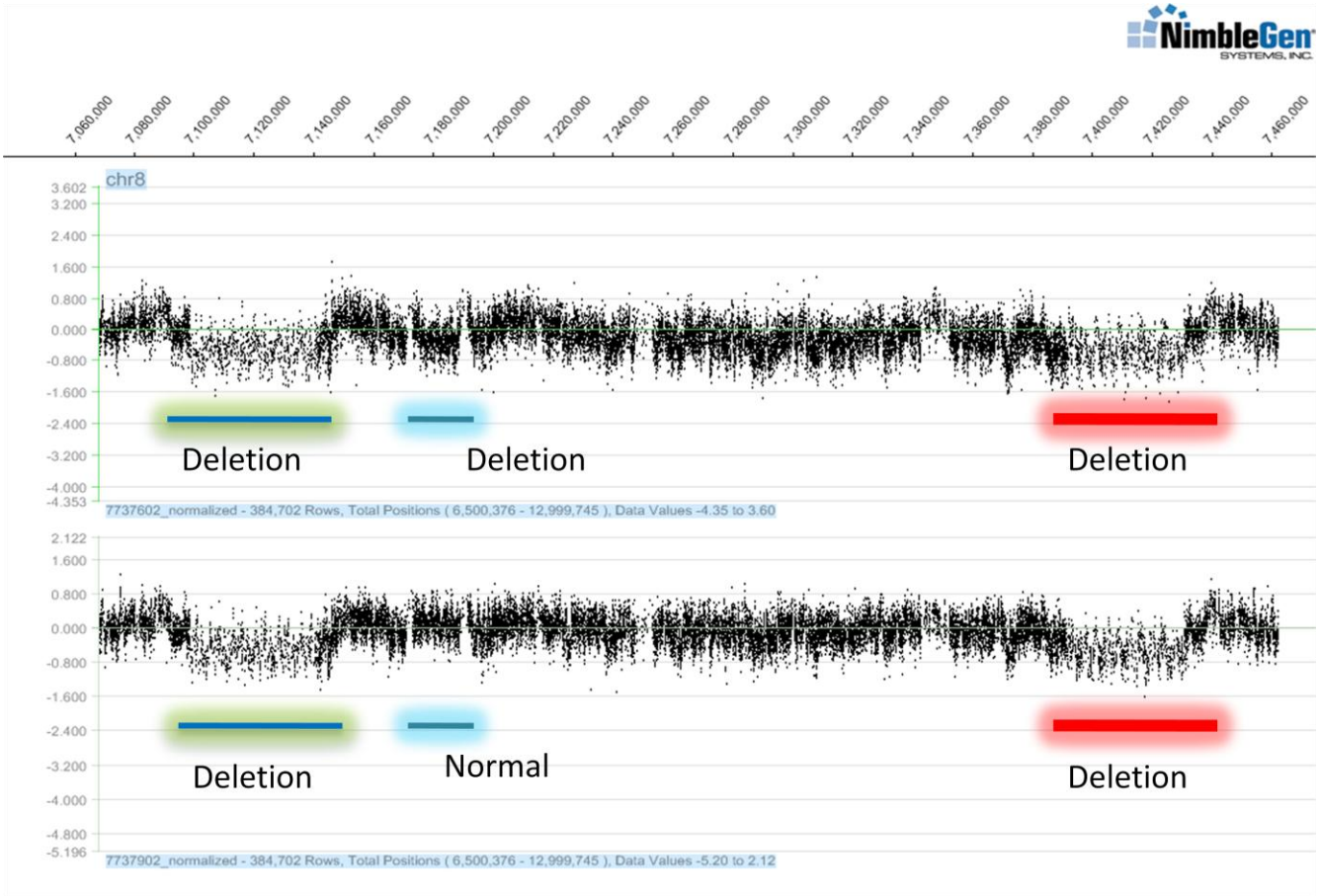


Figure 3.10 aCGH plot for the normalised data in two unaffected individuals for the region encompassing the *FAM90A* pseudogene cluster and beta-defensin gene cluster. Regions of reduced copy number can clearly be noted in the case of the green and red bars, with a distinct difference in copy number between the two samples being observed for the blue bar.

3.3.2 MeV Analysis of samples for CNVs

Analysis of the 100bp dataset revealed no regions that were either duplicated or deleted in all 5 KWE affected individuals (Table 3.5). Two deleted regions were identified in three of the five affected individuals, however these regions are found within the distal olfactory repeat region. The analysis of the complete normalized dataset revealed many regions that showed copy number variation only in the affected individuals, however no common regions were identified consistently in all 5 affected individuals. There were however several regions that showed deletions in three of the five affected individuals and one region that was deleted in four of the affected individuals.

The regions found to show copy number variation in three of the affected individuals were further investigated and many were found to occur within the beta-defensin gene cluster with two deletions occurring within *DEFB104A* and *DEFB107A*. The regions surrounding the various beta-defensin genes were subjected to BLAST searches against the human genomic reference sequence (HG18) and where found to occur at a number of locations on chromosome 8 within the repeat regions. These regions therefore share sequence homology with a number of loci on chromosome 8 and therefore the intensities observed at a set of probes could be due to the hybridization of sequences from a number of highly homologous loci. This made it impossible to distinguish the copy number of each of these loci individually. Increased copy number of the beta-defensin genes including *DEF104A* and *DEF107A* has however previously been shown to be associated with the skin disorder psoriasis(Hollox et al., 2008). In this case, however, there did not appear to be any consistent increase or decrease in copy number of specific beta-defensin genes in the affected individuals.

Copy number variation was also observed within the intronic regions of annotated genes, however not within any genes in the KWE critical region chr8:11,415,050-12,886,829. The critical region contains at least six well characterized genes including *BLK, GATA4, NEIL2, FDFT1, CTSB* and *LONRF1*. None of these genes showed any variation in copy number. Genes that were shown to possess some level of CNV within their coding sequence included *MSRA, RP1L1, PINX1* and *XKR6.* Although these genes were found outside of the KWE critical region, investigation of their molecular function did not suggest any link with the pathogenesis of KWE.

Table 3.5 Copy number variation observed in affected samples in both the normalized and 100bp averaged datasets using MeV.

Normalised Dataset						
Common Deleted Regions	**Samples**					**Closest Known Region**
7119740	2502	8002	9402			Downstream *DEFB103A* chr8:7273826-7275092
7131188	2502	8002	9402			Downstream *DEFB103A* chr8:7273826-7275092
7134548	2502	8002	9402			Downstream *DEFB103A* chr8:7273826-7275092
7141881	2502	8002	9402			Downstream *DEFB103A* chr8:7273826-7275092
7235589		8002	9402	5202		Downstream *DEFB103A* chr8:7273826-7275092
7267619		8002	9402		9302	Downstream *DEFB103A* chr8:7273826-7275092
7315664		8002		5202	9302	*DEFB104A/B chr8:7315241-7320014-Intronic region* [a]
7385662		8002	9402	5202		Downstream *DEFB107B* chr8:7340778-7354243
7396040	2502	8002	9402			Downstream *DEFB107B* chr8:7340778-7354243
7399547	2502	8002	9402			Downstream *DEFB107B* chr8:7340778-7354243
7666127	2502	8002	9402			Upstream *DEFB107A* chr8:7706652-7710648
7671765		8002	9402	5202		Downstream *DEFB107A* chr8:7706652-7710648
7707275		8002	9402	5202		*DEFB107A chr8:7706652-7710648-Intronic region* [a]
7813392		8002	9402		9302	Downstream *DEFB4* chr8:7789609-7791646
7816127		8002	9402	5202	9302	Downstream *DEFB4* chr8:7789609-7791646
7832001			9402	5202	9302	Downstream *DEFB4* chr8:7789609-7791646
8336792	2502	9302	9402			Upstream *PRAGMIN* chr8:8212668-8276667
10230336	2502	8002	9402			*MSRA chr8:9949240-10323809 -Intronic region* [a]
10365044	2502	8002	9402			Downstream *MSRA* chr8:9949240-10323809
10376988	2502	8002	9402			Downstream *MSRA* chr8:9949240-10323809
10562516	2502	8002	9402			*RP1L1 (AK127545) chr8:10,516,016-10,607,107-Intronic region* [a]
10685589	2502	8002	9402			*PINX1 (NM_017884) chr8:10660294-10734709- Intronic region* [a]
10839888	2502	8002	9402			*XKR6 chr8:10791067-11096285-Intronic region* [a]
10897727	2502	8002	9402			*XKR6 chr8:10791067-11096285-Intronic region* [a]
11168769	2502	9302	9402			Upstream *MTMR* chr8:11,179,410-11,223,062

12071685	**2502**	**8002**	**9402**			**Upstream *DUB3* chr8:12032086-12033678**[b]
12315046	2502	8002	9402			Upstream *DEFB130* chr8:12212843-12220196
12969776	2502	8002		5202		Downstream *DLC1* chr8:12985243-13416766
Common Amplified Regions	**Samples**					**Region**
6891692	2502	8002	9302			Downstream *DEFA5* chr8:6,900,239-6,901,669
9886092	2502	8002	9402			Downstream *TNKS* chr8:9450855-9677265
11082753	2502	8002	9402			*XKR6 chr8:10791067-11096285-Intronic region*[a]
11771928	**2502**		**9402**		**9302**	**Upstream *CTSB* chr8:11737445-11763055**[b]
12219524	2502	8002	9402			*DEFB130* chr8:12212843-12220196
12219718	2502	8002				*DEFB130* chr8:12212843-12220196
100bp Averaged Dataset						
Common Deleted Regions		**Samples**				**Region**
7131200-7131300	2502	8002	9402			Downstream *DEFB103A* chr8:7273826-7275092
7405600-7406000	2502	8002	9402			Downstream *DEFB107B* chr8:7340778-7354243

[a] Regions within genes are shown in italics with white backgrounds.

[b] Copy number variants within the vicinity of KWE candidate genes

3.3.3 Two regions of copy number variation for further investigation

Two regions of interest that exhibited copy number variation were identified. The first lies upstream of *deubiquitinating enzyme 3 (DUB3)* (Figure 3.11). This region was found to be deleted in three of the five affected individuals and lies 38007bp upstream of *DUB3* and is 24bp in length. Although this region lies significantly upstream of *DUB3* and out of range of a promoter region, *DUB3* is a newly characterized gene which codes for a single exon and has no know 5' UTR and therefore may have regulating factors that lie significantly upstream of the gene. BLAST results revealed that the sequence does share similarity with a number of other regions on chromosome 8 and therefore the result may be a false positive due to cross hybridization. This region is still a possible candidate region for KWE due to its close proximity to the newly identified KWE candidate gene *DUB3.*

The second region lies 8873bp upstream of *CTSB* and was found to be duplicated in three of the affected individuals (Figure 3.11). *CTSB* is a strong candidate gene for KWE, however no mutations have been identified in the gene to date. This 70bp duplicated sequence appears to be unique and therefore it is likely to be a real copy number variant and could possibly be involved in the transcriptional regulation of *CTSB.*

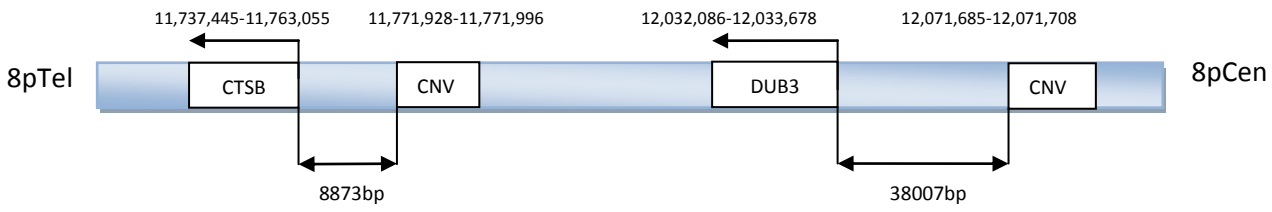


Figure 3.11 Two identified copy number variants and their locations in relation to KWE genes of interest

Chapter 4. Discussion

4.1 Discussion outline

The findings of this study aimed to provide further insight into the KWE linked critical region through the identification and annotation of novel functional elements in the region. This was achieved via two main approaches; a comparative genomic analysis of the KWE linked critical region and an analysis of the KWE linked critical region for copy number variation. To date the molecular cause of KWE remains unknown although the disease was first identified over 3 decades ago(Findlay et al., 1977) with the first linkage studies starting in 1985 and the gene been localized to 8p in 1997 (Starfield et al., 1997). A number of positional candidate genes have been examined for mutations in affected individuals, however none has presented with significant evidence to be linked definitively to KWE. The elusive nature of the associated genetic basis for KWE suggests that the molecular cause may lie in a previously uncharacterised element within the critical region. The results of this study aimed to address this possibility.

The comparative genomic analysis of the KWE linked critical region led to the discovery of a novel gene structure, *DUB3,* which upon further investigation presented as a very strong functional candidate gene for KWE. The approach, as demonstrated in previous studies, proved successful in identifying functional regions in the genome based on the conservation of these regions in a range of evolutionarily diverse organisms. In addition, a number of possible regulatory regions were identified in the form of putative transcription factor binding sites, and possible non-coding RNA genes although these were not able to be confirmed with much confidence based on computational predictions.

The aCGH of the KWE critical region led to the identification of a number of copy number variable regions within the KWE critical region. Variation in the copy number of functional segments of the genome has been implicated as the cause of a number of inherited disorders (Stankiewicz and Lupski). No consistent copy number variation was identified in the

KWE affected individuals, suggesting that the molecular etiology of KWE might not be due to alterations in copy number of functional segments in the KWE critical region.

Since the aims of this study were divided into two main components the structure of the following discussion will be divided into two sections. The first will look at the identification of novel functional elements in the KWE critical region and explore the possible link that these regions might have to the KWE phenotype, followed by the identification and analysis of copy number variants observed within the KWE critical region.

4.2 Comparative genomics for the identification of functional elements in the human genome

The comparative genomic analysis of the KWE critical region identified 8 regions that displayed high levels of sequence conservation across the 8 evolutionarily diverse organisms. The inclusion of a range of evolutionarily distant organisms such as the chicken, zebrafish, pufferfish and frog was aimed at being able to identify deep phylogenetically conserved regions. The addition of more closely related organisms such as the primates and mammals was aimed at the identification of highly constrained functional elements. The comparison of genomic sequences at different divergence levels has proven more reliable in the identification of regulatory sequences which unlike protein-coding regions are known to be under varying levels of constraint (King et al., 2007). Deciphering the relationship between function and conservation of non-coding regions is still a topic of much research in recent years which has been spurred on by the rapid generation and availability of whole genome sequences from a variety of different organisms.

Although the area of comparative genomics has matured significantly in the last 5 years, the challenges involved in accurately identifying regions of the genome that are under functional constraint to remain unchanged, are far from being concluded. The underlying premise that forms the basis of comparative genomics has however been well established since the first comparison of genomic sequences was conducted following the sequencing of the human, mouse and rat genomes. The initial comparison of the mouse and human genomes revealed

that almost 40% of the mouse genome aligned to the human genome with an average nucleotide identity of 67.2% (Waterston et al., 2002; Hardison, 2003). An estimated 5% of the 40% of aligned sequence encompasses protein-coding genes and associated regions shared between the two organisms which are under strong selective pressure to remain unchanged. The function of the remaining 35%, although highly conserved, remains unknown. It has been estimated that the rate of divergence in independently evolving vertebrate genome is on average 0.1-0.5% per million years (Tautz, 2000). The estimated last common ancestor shared between mice and humans approximately 80 million years ago placed them at a strategic position to identify shared functionally conserved sequences. Subsequently a number of additional mouse-human comparisons validated this finding through the identification of a number of additional functional sequences based solely on this pairwise analysis (Loots et al., 2000; Pennacchio and Rubin, 2001; Gottgens et al., 2002). These findings led to the development of the application of the genomic comparison of sequences to identify previously unannotated novel genes, larger functional regulatory regions as well as the identification and characterization of transcription factor binding sites in conserved non-coding sequences.

With the generation of a number of additional genomic sequences from a range of evolutionarily diverse organisms, the application of comparative genomics was extended rapidly to include the comparison of phylogenetically diverse organisms. These studies which explored the conservation of sequences across both mammalian and non-mammalian organisms demonstrated that in addition to functional protein coding-regions, a substantial fraction of non-coding sequence is also strongly conserved between evolutionarily diverse organisms (Waterston et al., 2002; Dermitzakis et al., 2003; Thomas et al., 2003; Dermitzakis et al., 2005). This led to the realization that the inclusion of a range of evolutionarily diverse organisms in comparative strategies was appropriate to identify both conserved slowly evolving functional protein-coding regions between closely related organisms as well as conserved non-coding regions of the genome which may harbour possible novel functional regions.

For the reasons mentioned above the comparison of the KWE critical region included a range of evolutionarily diverse organisms ranging from the closely related primates to the evolutionarily divergent chicken, zebrafish, pufferfish and frog. It must however be noted that the choice of species in a comparative sequence analysis is a challenge and in most instances is derived based on available data and the selection of organisms that share a common physiology or biology. In most studies this approach is useful, however in instances where the comparison is between closely related organisms, the conservation is usually too high leading to the masking of functionally relevant regions. This was clearly observed in the conservation plots (Appendix B) for the KWE critical region between the closely related primate lineages. The comparison of closely related organisms has proven useful in the identification of functional elements sharing a common biological role in the organisms compared. However, a balance of biological relevance, evolutionary distance and sequence analysis, is the most suited approach for the identification of conserved sequences that appear to be under evolutionary constraint against the background of sequence that has randomly diverged due to genetic drift (Nobrega and Pennacchio, 2004).

4.2.1 Identification of evolutionarily conserved elements

In the present study 8 regions were identified that exhibited high levels of sequence conservation across the KWE critical region. The 8 regions were found to exhibit some level of evolutionary constraint based on the fact that they displayed high levels of sequence similarity and were therefore subject to some sort of evolutionary pressure to remain unchanged. As previously mentioned for the purpose of this study, regions of interest were those that displayed a sequence conservation of at least 70% and a minimum length of 100bp across at least 6 of the evolutionarily diverse organisms.

As expected, the annotated protein-coding genes within the region displayed a high level of sequence conservation within exonic regions. This was particularly observed for the two previous KWE candidate genes in the region *FDFT1* and *CTSB* that showed high levels of sequence conservation for exons down to the evolutionarily divergent zebrafish, pufferfish and frog (Figure 3.3) , suggestive of a shared biological role for these genes in all the

organisms studied. *FDFT1* was previously identified as one of the strong biological candidate genes for KWE and encodes a membrane-associated enzyme which is involved in the mevalonate pathway, a pathway responsible for the production of critical biological end products such as cholesterol, dolichol, ubiquinone, steroid hormones and prenylated proteins (Pandit et al., 2000). The enzyme is also essential for the production of essential lipids and additional proteins that are required for the regulation of keratinocyte differentiation and proliferation in the skin. Similarly *CTSB* is known to encode a protease that is believed to be involved in the control of epidermal keratinocyte proliferation, differentiation and migration which are essential processes involved in epidermal differentiation and keratinisation as well as involved in the repair mechanism of disrupted epidermal barrier (Buth et al., 2004). The strong evolutionary constraint of these regions highlights their important functional roles in all the species studied, however a recent study examining the genomic DNA and expression of these two candidate genes in KWE affected individuals failed to identify any variants that segregated with the disease and thus excluded them as likely candidate genes for KWE (Hobbs *et al*, in press).

The level of conservation observed between the two non-human primate lineages (chimp and rhesus) and human was significantly high across the KWE critical region. This finding was to be expected due the relatively recent divergence of the lineages approximately 6-7 million years ago. Although interesting events have taken place during this period which has led to changes in the human genome responsible for significant alterations in morphology, physiology and behavior in humans, the extremely high levels of conservation shared between humans and primates make it exceedingly challenging to identify conserved regions that are under positive pressure to remain unchanged due to functionality. The question of conservation versus constraint is raised when comparing sequences from two very closely related organisms.

4.2.1.1 Conservation versus constraint

The premise of the approach used in this part of the study was based on the underlying assumption that functional DNA sequences between the species examined have changed significantly less than neutral DNA over the relevant phylogenetic distances between them. Most evolutionary change between species occurs as a result of mutations with minimal or no functional impact that are fixed via random genetic drift. In contrast, mutations in functional elements such as coding exons or *cis*-regulatory elements are likely to impair function, be deleterious to the organism, and are subsequently eliminated by purifying selection. The detection of sequences affected by purifying selection, which are under evolutionary constraint, can be used to annotate novel functional sites in the genome. It is however important to distinguish between conservation and constraint. Conservation is simply the observation of similarity between sequences whereas constraint is the hypothesis about the effects of purifying selection. The conservation of a particular sequence when observed to be in excess of levels predicted by a neutral model, can be used to infer evolutionary constraint (King et al., 2007; Cooper and Brown, 2008). However, the presence of conservation does not necessarily imply constraint nor does its absence imply a lack of constraint. It remains a challenge to measure the effects of selection on any given sequence of the human genome when comparing only closely related genome sequences, such as non-human primates, as due to the recent divergence of these organisms a vast majority of neutral sequences remains conserved between these organisms.

The use of multi-species sequence alignments has been used to address this question and several studies have shown that many stringently constrained non-coding sequences are indeed functional, even across evolutionarily diverse organisms such as mammals and fish (Nobrega and Pennacchio, 2004; Woolfe et al., 2005; Bejerano et al., 2006). The comparison of distantly related organisms like the zebrafish and *Fugu* are so divergent that neutral sites have been completely saturated with changes and therefore any sequence that aligns between these organisms is almost certainly under constraint. In contrast, the converse also however holds true in the fact that some apparently constrained sequences may have little

or no obvious function. This has been observed in studies conducted by several groups on the globin gene complexes, where a number of cis-regulatory modules (CRMs) have been experimentally identified within the globin gene complexes however subsequent multiple sequence alignments revealed that some of them are not conserved between human and mouse alignments (Dermitzakis et al., 2002; Hughes et al., 2005; King et al., 2005). It should therefore be noted that the focus only on deep evolutionarily conserved regions in this study served as a means to hone in on regions that were most likely to represent sequences that were highly conserved due to constraint.

4.2.2 Identification of a novel gene *deubiquitinating enzyme 3* (*DUB3*)

The functional analysis of the conserved region 6 identified within the KWE critical region revealed a possible single-exon gene structure based on the computational predictions using both GENESCAN and GeneID. Coupled with the high levels of sequence conservation of the region, this provided a sufficient amount of evidence to suggest that the region was under selection and thus subjected to some level of functional constraint. Further evidence for the active transcription of the region was observed based on the mapping of a number of experimental mRNA sequences to the region. Single exon genes (SEGs) are characteristic of prokaryotic genome architecture due to the lack of the intron-exon gene structure observed in eukaryotic genomes; however several SEGs have been identified in eukaryotic genomes (Brosius, 1999; Gentles and Karlin, 1999; Venter et al., 2001; Sakharkar et al., 2004). The presence of SEGs in eukaryotic genomes is quite unexpected and yet to be completely understood, however the evolution and origin of some of these SEGs is thought to be due to retro-transposition of spliced mRNA derivatives of eukaryotic genes within the genome and accounts for approximately 3000 genes in the genome (Sakharkar et al., 2004). The prediction of SEGs has therefore in recent years been included in computational prediction programs such as GENESCAN and GeneID for the prediction of human gene structures.

4.2.2.1 Genome complexities of the DUB3 region

Prior to the start of the current study, the region had been defined as non-coding DNA although the gene *DUB3* had previously been characterised in a study by Burrows et al. in 2004 (Burrows et al., 2004). The gene had been identified as belonging to a group of deubiquitinating enzymes, which are a group of proteases responsible for the cleavage of ubiquitin from ubiquitin-conjugated proteins. To date five families consisting of approximately 95 members have been identified including the Ubiquitin C-terminal Hydrolases(UCHs), the Ubiquitin Specific Proteases (USPs), the Machado-Joseph Disease Protein Domain Proteases (MJDs), the Ovarian tumour Proteases (OTUs) and the JAMM Motif Proteases (Wilkinson, 1997; Komander et al., 2009). *DUB3* belongs to the DUB/USP17 subfamily which is the largest family, consisting of approximately 56 members. These proteins are characterised by their two well conserved motifs known as the Cys and His boxes which are essential for their catalytic activity (Wilkinson, 1997). The DUB3/USP17 subfamily of deubiquitinating enzymes was first identified as immediate early genes induced in response to cytokine stimulation in both mice (*DUB1, DUB1A, DUB2*) and humans (*DUB3*) (Zhu et al., 1996b; Zhu et al., 1997; Baek et al., 2004; Burrows et al., 2004).

Previous studies suggested that the murine genes originally identified were part of a head-to-tail repeat of DUB/USP17 genes on chromosome 7 of the mouse which had arisen due to a tandem duplication event (Zhu et al., 1997). Initial studies also reported that the human DUB3/USP17 is encoded as a 1593bp ORF within a megasatellite tandemly repeated region on chromosome 4p15 (Kogi et al., 1997) which was shown to be highly polymorphic ranging in copy number from 20 to 103 (Okada et al., 2002). In addition, it was also demonstrated that a limited number of repeat units are present on chromosome 8(Gondo et al., 1998). Although these repeat units were identified on chromosome 8 previously, due to the initial mapping of *DUB3* to chromosome 4, the region remained annotated as non-coding on chromosome 8 at the beginning of this study.

Furthermore recent evolutionary studies of the *DUB3* genes on both chromosomes 4 and chromosome 8 have revealed that there is a cluster of 3 *DUB/USP17* repeat units on

chromosome 8 which appears to have arisen as a result of a duplication of one of the entire repeat unit blocks as opposed to a tandem duplication of individual sequences as was found on chromosome 4 (Burrows et al., 2010). These clusters occur at varying positions on chromosome 8 with only the annotated functional *DUB3* occurring within the KWE critical region. This hypothesis of a duplication of an entire block of sequence encompassing the *DUB3* gene was confirmed by analysis of the surrounding sequence.

It has previously been shown that a beta-defensin gene cluster also occurs on chromosome 8 (Hollox, 2008). The cluster consists of *DEFB107, DEFB105, DEFB106, DEFB104, SPAG11, DEFB103, DEFB4, DEFB108* and *DEFB109* surrounded by *FAM90A/*olfactory receptor genes which have been shown to be variable in copy number (Groth et al., 2008; Hollox, 2008). Two of the DUB/USP17 repeat units are present within the mapped beta-defensin gene cluster (Figure 4.1 A &B), with the remaining *DUB3* containing cluster not found within the recognised copy number variable beta-defensin cluster but associated with a cluster of other beta-defensins , *FAM90A* and olfactory receptor genes (Figure 4.1 C.) This suggests that although the same conserved blocks appear to be derived from a common ancestral block, the third region which contains the functionally annotated *DUB3* gene is not located within the copy number variable beta-defensin cluster.

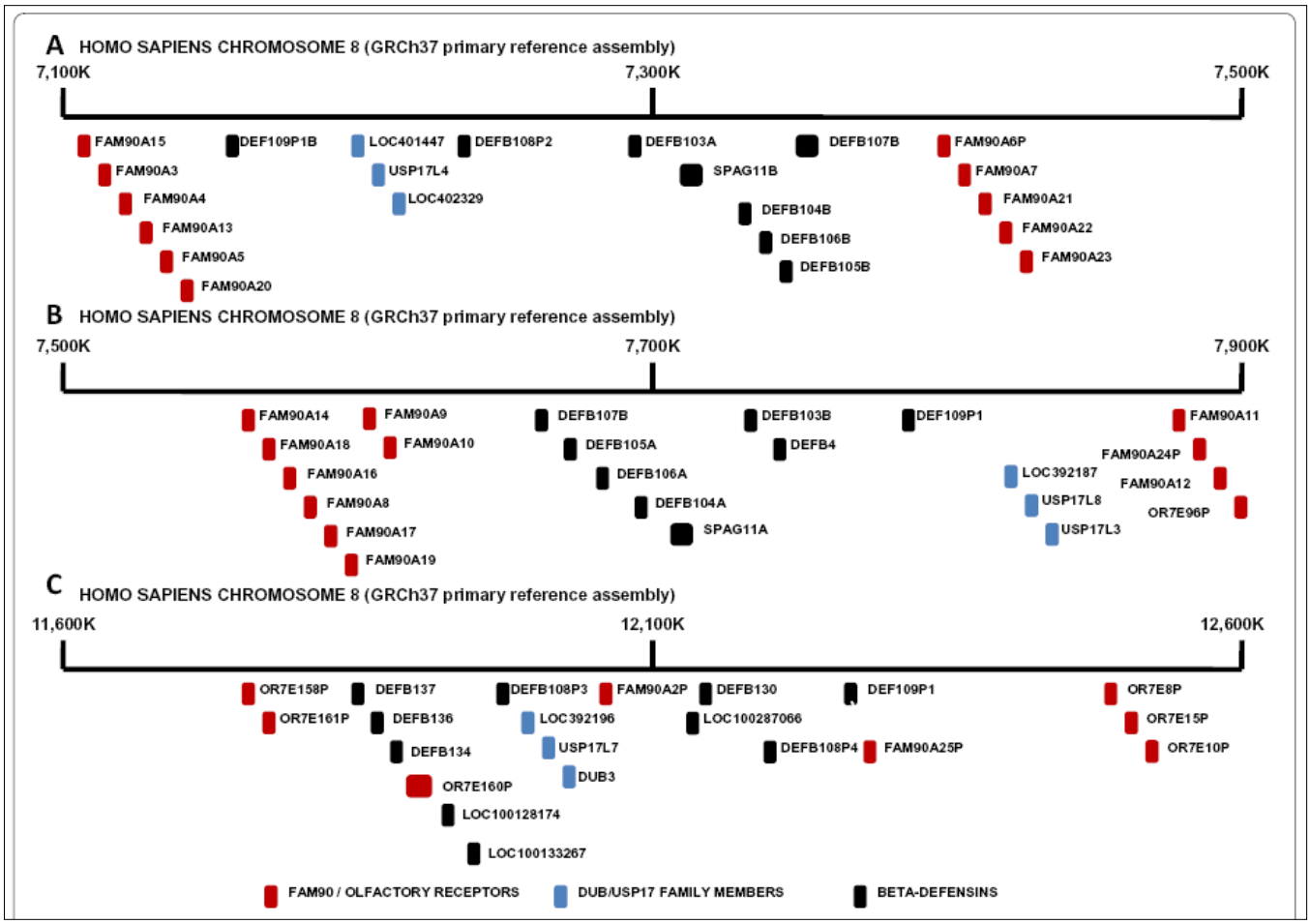


Figure 4.1 Regions on chromosome 8 surrounding DUB/USP17 family members. Diagrams of chromosome 8 (GRCh37 primary reference assembly) from bases (A) 7 185 000 – 7 205 000 (B) 7 820 000 – 7 840 000 and (C) 11 980 000 – 12 000 000. The approximate position of the identified DUB/USP17 sequences as well as the adjacent beta-defensin, FAM90A and olfactory receptor genes are indicated by the boxes illustrated in the key below (Burrows et al., 2010).

The above mentioned studies served to explain why *DUB3* had not previously been annotated to the KWE critical region. The focused analysis of the KWE critical region did however lead to the successful identification of the gene as a possible protein-coding region. To date there is no experimental evidence to suggest that any of the additional intact DUB/USP17 ORFs apart from *DUB3* on chromosome 8 and *USP17* on chromosome 4 are active.

4.2.2.2 DUB3 as a possible KWE candidate gene

DUB3 contains two conserved domains, UBP5 and Peptidase_C19E. These domains were identified through the alignment of the human DUB3 protein sequence with that of the mouse DUB1, DUB2 and DUB2A for which the functional domains have already been characterised (Burrows et al., 2004). UBP5 represents the ubiquitin C- terminal hydrolase which is believed to be responsible for the post translational modification, protein turnover and chaperone activity of the *DUB3*. Peptidase _C19E is part of the Peptidase C19 family which contains ubiquitinyl hydrolases which are intracellular peptidases that remove ubiquitin molecules from polyubiquitinated peptides by cleaving isopeptide bonds. The DUB enzymes act at several points in the ubiquitin pathway, which plays an important role in many molecular mechanisms such as cell division, cell cycle regulation , transcription, differentiation and development, apoptosis regulation and in the degradation of short lived regulatory proteins. In particular *DUB3* expression is regulated by the cytokines interleukin 4 (IL-4) and interleukin 6 (IL-6) (Burrows et al., 2004). Cytokines are polypeptide growth factors that are involved in transmitting signals between cells. Cytokines regulate immunity and inflammation and are produced in response to an immune stimulus and are known to be synthesized and secreted by keratinocytes (Feliciani et al., 1996). Keratinocytes represent the majority cell type in the skin and the production of cytokines from epidermal cells maintains the normal homeostatic mechanisms in the skin and can induce dramatic changes due to injury or any external stimulus (Feliciani et al., 1996). Cytokines may have both a pro-inflammatory and anti-inflammatory function and improper regulation of cytokines have been implicated in a number of skin diseases, including psoriasis, atopic dermatitis, lupus erythematosus and squamous cell carcinoma (Feliciani et al., 1996).

Further studies have implicated the *DUB* genes in cell growth and survival, in particular *DUB1* expression results in cell cycle arrest prior to the S-phase, while *DUB2* expression has been found to directly inhibit apoptosis induced by cytokine withdrawal (Zhu et al., 1996a; Migone et al., 2001). More recent studies have shown that the constitutive expression of *DUB3* can directly block cell proliferation via the regulation of the ubiquitination and activity of the

"CAAX" box protease Ras converting enzyme 1 (RCE1) (Burrows et al., 2004; Burrows et al., 2009). In addition, it has also been found that overexpression of the UPS17 family members can lead to apoptosis and in particular a recent study has shown that *DUB3* regulates the ubiquitination and stability of *CDC25A* which is known to play an integral role in the progression of the cell cycle.

DUB3 therefore presented as a strong candidate gene for KWE, not only because of its high level of conservation, but also due to the involvement in blocking cell proliferation and inducing apoptosis (Burrows et al., 2004). Apoptosis is a highly conserved basic physiological process in which dead unwanted cells are removed without inducing an inflammatory response. In self-renewing tissue like the epidermal layers of the skin, cell numbers are regulated by a delicate balance between proliferation, differentiation and cell death (Reefman et al., 2005). In some human skin diseases such as psoriasis, apoptosis is thought to play a role in the development of the disease (Chen et al., 2006). The induction of apoptosis results in an increase in the amount of dead skin cells which need to be removed from the skin's stratum corneum. If for some reason, the balance between proliferation, differentiation and apoptosis is disrupted, it could lead to the development of abnormalities. It has been demonstrated that the transient expression of *DUB3* is regulated by the cytokines interleukin 4 (IL-4) and interleukin 6 (IL-6). Any form of stress to the palmoplantar surfaces of the skin will result in the stimulation and release of cytokines, including IL-4 and IL-6, which will stimulate the transient expression of *DUB-3.* The constitutive expression of *DUB3* will block growth factor-dependent cell proliferation and possibly induce apoptosis. *DUB3* may function to regulate immune responses by influencing cell proliferation and survival (Burrows et al., 2004).

The comparative genomics approach therefore proved successful in the identification of not only a novel gene in the KWE critical region, but a gene which based on what is known about its molecular function presented as a strong functional candidate for the KWE phenotype. The gene was subsequently analysed by a colleague, Miss Angela Hobbs, in a sample of KWE affected individuals to identify possible mutations within the coding sequence of the gene.

Six genetic variants were identified within the *DUB3* genomic sequence, however all six variants were observed at a frequency of >70% in a corresponding control group analysed (Hobbs, Unpublished). Since KWE has displayed a strong autosomal dominant inheritance pattern and high penetrance these variants are unlikely to be the cause of KWE. However these findings do not completely exclude *DUB3* as a possible molecular cause for KWE since an investigation into the expression of the gene in KWE affected individuals is yet to be explored. Since *DUB3* is an excellent candidate gene for KWE it is possible that there may be variants present outside the coding region of the gene which may be involved in controlling the levels of expression of the gene. As previously mentioned, *DUB3* is constitutively expressed in response to external stimuli by cytokines, hence it is expected that there should be a number of *cis*-regulatory modules (CRMs) either close to or distant from the gene which are responsible for the regulation of the gene's expression. Variants within these CRMs which alter binding sites for necessary regulatory proteins such as transcription factors could directly affect the expression of the gene.

Although a complete analysis of the expression levels of *DUB3* in KWE affected individuals is required, there are a number of challenges in conducting this study effectively. Firstly, *DUB3* is a single-exon gene with no described 5' UTR, hence if varying levels of expression are observed in KWE affected individuals it will be challenging to determine exactly where the regulatory regions are located. Secondly, due to the high similarity between *USP17* located on chromosome 4p15 and *DUB3* it would be challenging to attempt to differentiate between the transcripts encoded by these two genes and hence accurately determine the expression levels of *DUB3.*

4.2.3 Transcription factor binding sites within conserved regions

The computational identification of TFBS within the human genome remains a challenging task. In the present study a contributing factor to the complexity of identifying novel TFBS was confounded by the fact that the conserved regions analysed were not found within close proximity to any known genes. Although most transcriptional regulation of genes are mediated by the sequence specific binding of a combination of TFs to TFBSs close to the

transcription start site, the formation of a transcription initiation complex (TIC) which is responsible for the recruitment of RNA polymerase involves the interaction of a number of additional elements which may not necessarily be found close to the gene that is being regulated (Maston et al., 2006). One of these elements may include distal TFBSs which may bind TFs a significant distance from the gene that they regulate and contribute to the formation of the TIC via protein-protein interactions or chromatin remodeling complexes together with non-DNA-binding co-factors and a cluster of TFBS which together form a CRM (Figure 4.2). These distal binding sites may be either enhancers or silencers of gene expression and ultimately control the spatial and temporal expression of genes. The conserved regions identified in this study could therefore contain TFBS which might be involved in the spatial and temporal regulation of genes in the KWE critical region.

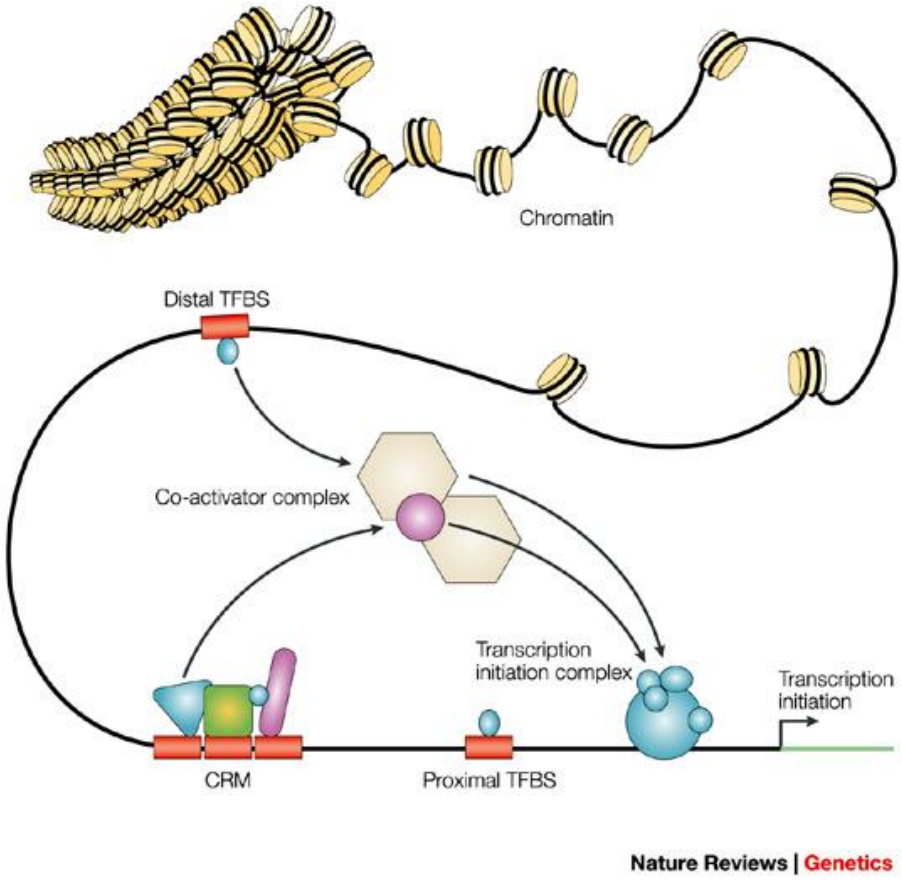


Figure 4.2 Diagrammatic representation of all the possible elements involved in the regulation of gene expression. Transcription factors can bind to specific sites that are either proximal or distal to a transcription start site. (Wasserman and Sandelin, 2004)

In the present study a number of transcription factor binding sites were identified within the conserved regions 3, 4, 5 and 8 using the conventional approach for the identification of motifs based on the position weight matrices (PWMs) for known TFs stored in the JASPAR database. As previously mentioned this approach treats each site independently and even though the PWMs within the JASPAR database contain only PWMs for experimentally validated TFs, the scanning of the conserved sequences with these PWMs led to a number of spurious matches which may occur in the genome by chance and not be functional *in vivo*. This was certainly apparent since although only the top scoring hits are shown in Table 3.3, the total number of predicted TFBS for regions 3, 4, 5 and 8 were 106, 42, 39 and 80 respectively. The addition of the phylogenetic footprinting approach was used to in order to address this problem of the overrepresentation of possible false positive TFBS. The

consistent prediction of a SOX9 binding site for region 3 provided some additional evidence that the site may indeed be a likely functional site. Similarly the consensus predictions using both methods for region 4 and 5 were that of an SP1 TFBS, while region 8 produced completely conflicting results using the two methods. These results served as an example of the varying predictive power of the different approaches for identifying *de novo* TFBS based purely on genomic sequences. An assessment of 13 publicly available methods for the prediction of transcription factor binding sites showed that no method surpassed others in predictive power of a set of data, indicating that the problem of motif discovery is far from solved (Tompa et al., 2005).

4.2.3.1 Computational predictions of regulatory elements

Ultimately the conservation of the regions above together with the prediction of putative TFBS presents convincing evidence to suggest that these regions may harbour some regulatory elements. Although the results focused on the highest scoring predictions, it is also possible that these conserved regions may harbour multiple TFBS, since as previously described; transcriptional regulation usually involves an interaction between a number of TFs, chromatin remodeling complexes and other non-DNA-binding co-factors. In essence activation or repression of a gene is seldom a binary switch whereby the gene is either switched on or off. It is rather a modulation of the rate of transcription through a complex process involving a number of TFs and TFBS. These predicted TFBS within the conserved regions could be involved in the complex regulation of any of the genes either within close proximity or significantly far away. Mutations in regulatory elements are generally assumed to not have a pronounced phenotypic effect, since the expression pattern as opposed to the structure or function of the protein is altered. In addition, as previously mentioned it has been found that a single gene may have multiple regulatory elements with each having a small contributory function (Nobrega et al., 2003; Pennacchio et al., 2006). Although this has been observed, a number of mutations within non-coding sequences close to or within genes have been noted and include diseases such as macular degeneration (Dewan et al., 2006), pyruvate kinase deficiency (Manco et al., 2000), Hirschsprung disease (Emison et al., 2005)

and X-linked adrenal hypoplasia (Goto and Katsumata, 2009). In addition, diseases such as pre-axial polydactyly have been found to be caused by a mutation within a well conserved region located approximately 1Mb away from the associated gene *SHH* (Lettice et al., 2002) *,* while a mutation within an enhancer located 10kb upstream of *IRF6* has been associated with cleft lip (Rahimov et al., 2008).

These regions could possibly harbour regulatory elements, which are involved in the regulation of genes in the KWE critical region and therefore present as possible candidate regions for further investigation. Unfortunately the computational approach to identifying regulatory elements possesses its limitations. Recently high-throughput methods such as chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) have been used to identify genomic sequences bound by a particular protein *in vivo*. These types of experiments allow for the identification of different types of regulatory elements *in vivo*. Such studies are currently being conducted on a genome-wide scale and will provide more experimentally based evidence for regulatory regions and insight into the regulation of the annotated genes within the KWE critical region.

4.2.4 Identification of RNA secondary structure

Regions 4, 5 and 8 did not produce any biologically relevant secondary structures. Although all three predicted secondary structures contained possible hairpin loops, which are characteristic of miRNAs, the thermodynamics of the structures did not provide strong evidence for the formation of the particular secondary structures. A larger proportion of mammalian miRNAs have been mapped to introns of protein coding genes or long mRNA-like non-coding transcripts (Rodriguez et al., 2004; Kim and Kim, 2007). In addition, increasing evidence suggests that miRNAs are regulated and transcribed similar to protein-coding genes (Soifer et al., 2007). Since regions 4, 5 and 8 all have experimental mRNA sequences mapped to them, it is possible that they may form part of larger transcribed region within the KWE

critical region which is yet to be annotated. It would however be expected that the conserved regions within a non-coding mRNA transcript would be the regions encompassing the functional components of the non-coding RNA, i.e. the regions that are conserved in order to form hairpin secondary structures required for functional miRNA processing. Since these were the regions that were identified and analysed for secondary structure in this study it is unlikely that the conserved regions 4, 5 and 8 contain any functional non-coding RNA genes.

4.3 Copy number variation and KWE

The present study identified a number of regions that exhibited varying levels of copy number variation within the KWE critical region. It must be noted that since it is known that KWE exhibits a strong autosomal dominant pattern of inheritance and is known to occur due to a founder effect, it was expected that if the molecular cause was due to copy number variation this too would be present as a common feature across all five of the affected individuals. Hence the premise for the analysis of the aCGH data was based on the fact that, if KWE was caused by a duplication or deletion, we expected to find a common pattern of CNV in all KWE affected individuals when compared to the healthy controls. However, no region was found which displayed a consistent CNV across all 5 affected individuals when compared to the controls. None the less, a number of regions of copy number variation were observed within regions with potential clinical links to the KWE phenotype which are discussed further below.

4.3.1. CNV of the beta-defensin gene cluster

The regions that exhibited the most variation in copy number in the present study were observed within the REPP and REPD clusters. Since these regions are known to be composed of low copy repeats (LCRs), which are fragments of DNA greater than 1kb in size and possess a sequence identify of greater than 90%, it was expected that they may be responsible for mediating nonallelic homologous recombination (NAHR) resulting in the duplication or reciprocal translocation of segments of the region. The region was therefore expected to show a significant amount of neutral CNV with no significant phenotypic effect. The upstream cluster consists of *DEFB107, DEFB105, DEFB106, DEFB104, SPAG11, DEFB103, DEFB4, DEFB108 and DEFB109* and is flanked by *FAM90A/olfactory receptor gene* cluster.

Interestingly the beta-defensin cluster mentioned above is found to vary in copy number commonly ranging between 2 and 7 copies in healthy individuals (Groth et al., 2008; Hollox, 2008). The variation in copy number of this region appears to be a neutral polymorphism; however recent studies have shown variation in copy number of distinct defensin genes in

the region. Linzmeier and Ganz showed that *DEFB4* and *DEFB104* varied in copy number from between 2 and 8 with a mode of 6 copies per diploid genome (Linzmeier and Ganz, 2005). Interestingly subsequent studies revealed that epithelial *DEFB4* expression and the concentration of the corresponding peptide is regulated by the cytokine interleukin- 1 (IL-1) and that the mRNA levels of *DEFB4* were proportional to *DEFB4* gene copy numbers (Hollox et al., 2003). This finding therefore made it clear that the copy number of a gene can directly affect the expression levels of that gene. Although the region encompassing this cluster of beta-defensin genes lies substantially outside of the defined KWE critical region on the inverted chromosome, the cluster is incorporated into the region that was investigated.

Interestingly a later study looking at the possible role of the beta-defensin gene cluster in the inflammatory skin disorder psoriasis found a significant association between higher copy numbers of the repeat unit encompassing the beta-defensin genes *DEFB4, SPAG11, DEFB103, DEFB104, DEFB105, DEFB106* and *DEFB107* (Hollox et al., 2008). Psoriasis is a common inflammatory skin disease with a strong genetic component as well as an environmental contribution. A number of loci and alleles have been identified that confer risk of the disease with the beta-defensin cluster on chromosome 8 being identified as one of the genetic components. The disease is characterised by red-scaling and elevated plaques, commonly found on the elbows, knees and trunk. Beta-defensins are small secreted antimicrobial peptides which are encoded by the *DEFB* genes which are found in three main clusters; two on chromosome 20 and one on chromosome 8p23.1 (Ganz, 2003). *DEFB1* encoding the protein hBD-1 and *DEFB103* encoding the protein hBD-2 are expressed constitutively in the skin and *DEFB4* encoding the protein hBD-2 can be induced by cytokines in cultured keratinocytes (Harder et al., 1997). Since beta-defensins possess cytokine-like properties, high levels of expression may be induced due to skin injury or some other environmental trigger. In the case of KWE it is known that severity of disease symptoms is increased in the colder winter months during periods of low humidity. Low humidity has been found to stimulate keratinocyte proliferation and amplification of the hyper proliferative response to barrier disruption resulting in inflammation (Denda et al., 1998). The inflammation response is mediated by the beta-defensins and therefore it is possible

that increased copy number of the beta-defensin genes described above could lead to an inappropriate inflammatory response due to the environmental stimuli of low humidity levels resulting in the clinical symptoms typical of KWE.

Although in this particular study it was not possible to determine the exact copy number of the beta-defensin genes, since it is known that the copy number of these genes can vary in normal individuals, the variation observed within the cluster suggests that the copy number varies significantly between individuals with KWE. Although this is the case, no consistent duplications or deletions of specific beta-defensin genes were identified in the 5 KWE affected individuals in the present study. At this point the results do suggest that CNVs within the beta-defensin cluster described above may be linked to the clinical severity of KWE, but not essentially its presence or absence, and presents as an interesting avenue to be explored in further studies.

4.3.2 Copy number variation of known genes in the region

None of the previously characterised genes in the region displayed any consistent levels of copy number variation across their entire coding length in the present study. In particular the two main candidate genes, *FDFT1* and *CTSB* were both found to be present in the normal 2 copies for all the individuals analysed. Smaller regions of CNV were however observed within intronic regions of *MSRA, RP1L1, PINX1* and *XKR6*. Although these regions of copy number variation may be significant to the analysis of the molecular function of the genes they failed to identify any link to the phenotypic manifestations of KWE. Previous studies implicating CNV as the cause of molecular disorders usually result from the change in copy number of dosage sensitive genes (Lupski, 1992; Lupski, 2009). This has been observed in Charcot-Marie-Tooth disease type 1 A (CMT1A) and hereditary neuropathy with liability to pressure palsies (HNPP) where the duplication or deletion of the gene *PMP22* results in these two diseases, respectively (Chance and Fischbeck, 1994; Reiter et al., 1996). It was therefore expected that, should some occurrence of CNV variation be responsible for causing the KWE phenotype, it would encompass an entire gene leading to a gene dosage effect. In addition, in this case the altered CNV of the gene should be consistently observed across all KWE

affected individuals. Since this was not the case it seemed unlikely that KWE was caused by a change in copy number of a dosage sensitive gene within the critical region.

4.3.3 CNVs within potential regions of interest

In the current study a number of CNVs were identified within non-coding sequences of the region analysed. Although the most common molecular mechanism for CNVs to cause a specific phenotype is through the alteration of copy number of a gene or multiple genes sensitive to the dosage effect, CNVs can also result in a position effect (Kleinjan and van Heyningen, 2005; Lupski, 2007). A majority of these CNVs were small regions varying in size from 20bp to 200bp within the vicinity of known genes in the region. Although these regions were significantly small and unlikely to disrupt coding genes, it is possible that they may harbour regulatory elements required for the control of spatial and temporal expression of genes close to or significantly distant from them. Such effects have been observed in diseases such as spastic paraplegia type 2 where a microduplication located downstream from the *PLP1* gene has been associated with the disease (Lee et al., 2006). Similarly Dathe et al. detected a 5.5kb duplication in a non-coding sequence located approximately 110kb downstream of the *BMP2* gene in chromosome 20p12.2 as the cause of brachydactyly type A2 in two families (Dathe et al., 2009). The duplicated region was found to contain highly conserved sequences suggestive of a long-range regulator of the gene.

Although the regions of CNV observed in the present study are much smaller in size there is still a possibility that they might harbour regulatory elements utilized by one of the genes in the region. Two of the most plausible CNVs possibly involved in a position effect were those identified upstream of the two strong KWE candidate genes *CTSB* and *DUB3*. The CNV upstream of *CTSB* is located only 8873bp upstream of the gene and was found to be duplicated in 3 of the five affected individuals with no variation been observed in the control samples. It is possible that this region may harbour some regulatory elements involved in the control of expression of either *CTSB* or another gene in the region. Similarly, the CNV of 24bp found 38007bp upstream of the newly identified gene *DUB3* which was deleted in 3 of the 5 affected individuals could possibly be involved in the regulation of *DUB3.* Interestingly these

two regions did not fall within any of the highly conserved areas identified in the comparative genomics approach of the current study. In addition, as mentioned above most disease causing CNVs involving regulatory regions encompass large regions of regulatory sequences ranging in size from kb to Mb. It is therefore possible that although the location of these two CNVs in relation to candidate genes in the region presents them as likely areas for further investigation, it is unlikely that these regions alone would be involved in the molecular etiology of KWE.

4. 4 Limitations and future studies

4.4.1 Limitations

The comparative genomics analysis of the KWE critical region served as an exploratory method to provide further insight into the region which has been strongly linked to the disease. Although the comparison of orthologous sequences has proven to be a successful approach for the identification of potentially novel functional elements in the human genome, all findings require laboratory validation since computational approaches are far from being completely accurate. Although the approach led to the identification of a number of regions which appear to be highly conserved it remains a challenge to determine conclusively the molecular forces driving this conservation. Due to the extensive evolutionary diversity of the organisms included in the comparative sequence analysis it is possible that some of these regions may be conserved due to shared functional relevance, however the high level of divergence such as between the human and the pufferfish can lead to the loss of important mammalian specific functional regions. The present study focused exclusively on the identification of regions conserved in the most evolutionarily distant organisms. A drawback of this approach is that there may be additional regions that are conserved between the more closely related mammals which may be involved in mammalian specific functions which have been lost from more divergent species. The current study focused exclusively on the identification of conserved sequences across phylogenetically distant organisms and hence additional functional non-coding sequences may be identified through the examination of conserved regions present only in closely related organisms.

Although the array CGH data provided significant insight into copy number variation present within an extended area of the KWE critical region the method has a number of drawbacks which hinder the interpretation of results. Firstly, as previously mentioned in the methodology section each sample was hybridized to the array together with a control sample consisting of a pool of DNA from six unrelated Caucasian individuals. The downstream consequences of this are that there may be significant population differences in the copy number variation in the patients who are of South African Afrikaans origin that may

be masked due to allelic CNVs in the group of controls. Although the pooled control aims to overcome this factor, if there is a significant population difference in CNVs, as has been observed for SNPs, the copy number at known CNVs such as the beta-defensin gene cluster would be inaccurately represented. An attempt was made to address this problem through the pairwise comparison of affected and unaffected individuals within each family; however this did not yield any interesting results which pointed to a specific CNV allele consistently present in affected individuals.

Secondly, KWE has an autosomal dominant inheritance pattern with high penetrance. If KWE was caused by a CNV, it would be expected that the CNV would occur on one allele at a single locus. Although array CGH approaches are able to detect copy number variation at a single locus the resulting intensity observed is the "sum" of copy number at the two alleles for a particular individual. This may result in a masking effect as for a particular locus one KWE affected individual may possess three copies of a region on one allele and one on the other, while another affected individual may possess two copies on each allele resulting in the same intensity on the array. If the three copies were characteristic of a KWE allele it would be impossible to detect this based solely on the array CGH results.

4.4.2 Future Studies

Although the critical region analysed in this study has shown strong linkage to KWE, a thorough examination of the exact defining boundaries of the critical region is required via the typing of additional markers in affected and unaffected individuals. As mentioned in the limitations above, a more in depth exploration of evolutionarily conserved regions specifically in closely related organisms might lead to the identification of further functional regions that are yet to be annotated in the region.

The current advances in high throughput technologies in the field of molecular biology present a number of options for future studies in KWE. An examination of the gene expression patterns within KWE affected skin versus normal skin using microarray technology could provide some interesting insight not only into the molecular cause of the disease but also into the pathways involved in the modulation of the symptoms. The gene expression array should ideally include all of the characterised genes within the KWE critical region including the strong positional candidate genes examined in previous studies in order to understand if they are likely to be involved in the molecular etiology of KWE.

One of the recent advances in the field of disease gene discovery has been the development of an exome sequencing approach. Essentially this technology is based on a targeted array capture system followed by massively parallel sequencing which provides a genome wide approach to sequencing all expressed regions, including protein-coding regions (“exomes”) of the genome. Such a study would involve sequencing the “exomes” of a sample of KWE affected and unaffected individuals and indentifying variants within the exons present within the KWE critical region. Although the sequencing of several candidate genes in the critical region has already been done, this approach would allow for all genes within the KWE critical region to be assessed for variations regardless of their prior functional links to the KWE phenotype.

Finally, the most comprehensive approach for future KWE studies would be the complete sequencing of the KWE critical region, which has become a feasible option in recent years

due to a significant reduction in costs. The complete sequencing of the critical region in a significant number of affected individuals and healthy controls would allow for all variants in the region to be identified and their possible link to the KWE phenotype assessed. In contrast to the exome sequencing approach this method will allow for the identification of non-coding variants in addition to expressed and protein-coding variants.

4.5 Conclusions

The findings of the current study have provided a significant amount of further insight into the KWE critical region. The comparative genomic analyses of the critical region successfully identified a number of highly conserved regions which may be potentially unannotated functional regions. Further investigation of these regions led to the identification of a previously unannotated gene *DUB3* which presented as an excellent functional candidate for KWE. Although no possible pathogenic mutations were identified in the gene upon investigation, *DUB3* cannot conclusively be excluded as being linked to KWE until a comprehensive analysis of the expression pattern of the gene is conducted. The remaining conserved regions identified showed some evidence for the presence of regulatory elements in the form of TFBS, albeit solely based on computational predictions. These regions do however appear to be under some level of constraint to remain conserved across evolutionarily diverse organisms and therefore warrant some further experimental investigation to determine if they have potential functional roles. As further studies aimed at the genome wide identification of regulatory elements such as the ENCODE project and 1000 Genomes project are completed further insight will be gained into the correlation between computationally predicted TFBS and experimentally validated sites.

The copy number variation results suggest that the molecular cause of KWE is unlikely to be due to a variation in copy number of segments within the KWE critical region. It was expected that if this was the case the copy number variation would encompass a functional gene and this would be clearly observed in all of the affected individuals since KWE has a strong autosomal dominant inheritance pattern and is highly penetrant. However, as previously noted, the variation observed within the beta-defensin cluster did present as a possible avenue for further investigation since it has been previously implicated in skin disorders such as psoriasis. The determination of the exact copy number of these regions, which could not be determined in the present study, would be of interest. Although it may not be the sole molecular cause of KWE the copy number of the beta-defensin gene cluster may be involved in the interesting cyclic modulation of symptoms observed in KWE. The

absence of any consistent large copy number variants within all 5 of the affected individuals suggests that KWE is unlikely to be caused by a copy number variant as most other documented disease-causing CNVs are significantly larger than the regions of interest identified in the present study.

In conclusion, the comparative genomics approach employed in the present study proved successful in identifying a novel protein-coding gene *DUB3* which lies within the KWE critical region. *DUB3* has been identified as an excellent functional candidate gene for KWE although initial analysis of the gene in affected individuals revealed no possible pathogenic mutations. *DUB3* still however presents as an interesting candidate gene for further exploration. Additional putative functionally relevant highly conserved regions were identified within the critical region which was computationally predicted to harbour regulatory TFBS. The positioning of these regions within close to proximity to candidate genes necessitates the need for further experimental validation. The molecular cause of KWE is unlikely to be due to copy number variation within the KWE critical region, although further investigation of the beta-defensin cluster might yield some interesting results with regard to symptom modulation. The present study has provided valuable insight into the KWE critical region and revealed a number of potential regions of interests to be examined in further studies as the search for the elusive causative KWE gene continues.

References

Abbasi AA, Paparidis Z, Malik S, Goode DK, Callaway H, Elgar G, Grzeschik KH (2007) Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. PLoS ONE 2:e366

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LSt, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 310:317-320

Almeida AM, Murakami Y, Layton DM, Hillmen P, Sellick GS, Maeda Y, Richards S, Patterson S, Kotsianidis I, Mollica L, Crawford DH, Baker A, Ferguson M, Roberts I, Houlston R, Kinoshita T, Karadimitris A (2006) Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nat Med 12:846-851

Ambros V (2004) The functions of animal microRNAs. Nature 431:350-355

An P, Martin MP, Nelson GW, Carrington M, Smith MW, Gong K, Vlahov D, O'Brien SJ, Winkler CA (2000) Influence of CCR5 promoter haplotypes on AIDS progression in African-Americans. Aids 14:2117-2122

Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A 90:11995-11999

Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet 7:277-282

Appel S, Filter M, Reis A, Hennies HC, Bergheim A, Ogilvie E, Arndt S, Simmons A, Lovett M, Hide W, Ramsay M, Reichwald K, Zimmermann W, Rosenthal A (2002) Physical and transcriptional map of the critical region for keratolytic winter erythema (KWE) on chromosome 8p22-p23 between D8S550 and D8S1759. Eur J Hum Genet 10:17-25

Badano JL, Katsanis N (2002) Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet 3:779-789

Baek KH, Kim MS, Kim YS, Shin JM, Choi HK (2004) DUB-1A, a novel deubiquitinating enzyme subfamily member, is polyubiquitinated and cytokine-inducible in B-lymphocytes. J Biol Chem 279:2368-2376

Balikova I, Martens K, Melotte C, Amyere M, Van Vooren S, Moreau Y, Vetrie D, Fiegler H, Carter NP, Liehr T, Vikkula M, Matthijs G, Fryns JP, Casteels I, Devriendt K, Vermeesch JR (2008) Autosomal-dominant microtia linked to five tandem copies of a copy-number-variable region at chromosome 4p16. Am J Hum Genet 82:181-187

Barber JC, Maloney V, Hollox EJ, Stuke-Sontheimer A, du Bois G, Daumiller E, Klein-Vogler U, Dufke A, Armour JA, Liehr T (2005) Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. Eur J Hum Genet 13:1131-1136

Barber JC, Maloney VK, Huang S, Bunyan DJ, Cresswell L, Kinning E, Benson A, Cheetham T, Wyllie J, Lynch SA, Zwolinski S, Prescott L, Crow Y, Morgan R, Hobson E (2008) 8p23.1 duplication syndrome; a novel genomic condition with unexpected complexity revealed by array CGH. Eur J Hum Genet 16:18-27

Barrandon Y, Green H (1987) Cell migration is essential for sustained growth of keratinocyte colonies: the roles of transforming growth factor-alpha and epidermal growth factor. Cell 50:1131-1137

Beetz C, Schule R, Deconinck T, Tran-Viet KN, Zhu H, Kremer BP, Frints SG, et al. (2008) REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. Brain 131:1078-1086

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87-90

Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209-213

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394

Bonifas JM, Rothman AL, Epstein EH, Jr. (1991) Epidermolysis bullosa simplex: evidence in two families for keratin gene abnormalities. Science 254:1202-1205

Brosius J (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107:209-238

Burrows JF, Kelvin AA, McFarlane C, Burden RE, McGrattan MJ, De la Vega M, Govender U, Quinn DJ, Dib K, Gadina M, Scott CJ, Johnston JA (2009) USP17 regulates Ras activation and cell proliferation by blocking RCE1 activity. J Biol Chem 284:9587-9595

Burrows JF, McGrattan MJ, Rascle A, Humbert M, Baek KH, Johnston JA (2004) DUB-3, a cytokine-inducible deubiquitinating enzyme that blocks proliferation. J Biol Chem 279:13993-14000

Burrows JF, Scott CJ, Johnston JA (2010) The DUB/USP17 deubiquitinating enzymes: a gene family within a tandemly repeated sequence, is also embedded within the copy number variable beta-defensin cluster. BMC Genomics 11:250

Buth H, Wolters B, Hartwig B, Meier-Bornheim R, Veith H, Hansen M, Sommerhoff CP, Schaschke N, Machleidt W, Fusenig NE, Boukamp P, Brix K (2004) HaCaT keratinocytes secrete lysosomal cysteine proteinases during migration. Eur J Cell Biol 83:781-795

Calin GA, Croce CM (2007) Investigation of microRNA alterations in leukemias and lymphomas. Methods Enzymol 427:193-213

Chance PF, Fischbeck KH (1994) Molecular genetics of Charcot-Marie-Tooth disease and related neuropathies. Hum Mol Genet 3 Spec No:1503-1507

Chen X, Tan Z, Yue Q, Liu H, Liu Z, Li J (2006) The expression of interleukin-23 (p19/p40) and inteleukin-12 (p35/p40) in psoriasis skin. J Huazhong Univ Sci Technolog Med Sci 26:750-752

Collin C, Moll R, Kubicka S, Ouhayoun JP, Franke WW (1992) Characterization of human cytokeratin 2, an epidermal cytoskeletal protein synthesized late during differentiation. Exp Cell Res 202:132-141

Cooper GM, Brown CD (2008) Qualifying the relationship between sequence conservation and molecular function. Genome Res 18:201-205

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901-913

Costa JL, Meijer G, Ylstra B, Caldas C (2008) Array comparative genomic hybridization copy number profiling: a new tool for translational research in solid malignancies. Semin Radiat Oncol 18:98-104

Coulombe PA, Hutton ME, Vassar R, Fuchs E (1991) A function for keratins and a common thread among different types of epidermolysis bullosa simplex diseases. J Cell Biol 115:1661-1674

Dathe K, Kjaer KW, Brehm A, Meinecke P, Nurnberg P, Neto JC, Brunoni D, Tommerup N, Ott CE, Klopocki E, Seemann P, Mundlos S (2009) Duplications involving a conserved regulatory element downstream of BMP2 are associated with brachydactyly type A2. Am J Hum Genet 84:483-492

Davis E, Caiment F, Tordoir X, Cavaille J, Ferguson-Smith A, Cockett N, Georges M, Charlier C (2005) RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. Curr Biol 15:743-749

de Wilde J, Wilting SM, Meijer CJ, van de Wiel MA, Ylstra B, Snijders PJ, Steenbergen RD (2008) Gene expression profiling to identify markers associated with deregulated hTERT in HPV-transformed keratinocytes and cervical cancer. Int J Cancer 122:877-888

Denda M, Sato J, Tsuchiya T, Elias PM, Feingold KR (1998) Low humidity stimulates epidermal DNA synthesis and amplifies the hyperproliferative response to barrier disruption: implication for seasonal exacerbations of inflammatory dermatoses. J Invest Dermatol 111:873-878

Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet 6:151-157

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 420:578-582

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302:1033-1035

DeSilva U, Elnitski L, Idol JR, Doyle JL, Gan W, Thomas JW, Schwartz S, Dietrich NL, Beckstrom-Sternberg SM, McDowell JC, Blakesley RW, Bouffard GG, Thomas PJ, Touchman JW, Miller W, Green ED (2002) Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. Genome Res 12:3-15

Desnick RJ, Glass IA, Xu W, Solis C, Astrin KH (1998) Molecular genetics of congenital erythropoietic porphyria. Semin Liver Dis 18:77-84

Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 314:989-992

Di Nardo A, Wertz P, Giannetti A, Seidenari S (1998) Ceramide and cholesterol composition of the skin of patients with atopic dermatitis. Acta Derm Venereol 78:27-30

Dietz HC, Cutting GR, Pyeritz RE, Maslen CL, Sakai LY, Corson GM, Puffenberger EG, Hamosh A, Nanthakumar EJ, Curristin SM, et al. (1991) Marfan syndrome caused by a recurrent de novo missense mutation in the fibrillin gene. Nature 352:337-339

Donfack J, Schneider DH, Tan Z, Kurz T, Dubchak I, Frazer KA, Ober C (2005) Variation in conserved non-coding sequences on chromosome 5q and susceptibility to asthma and atopy. Respir Res 6:145

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223-227

Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, Frazer KA (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res 10:1304-1306

Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2:919-929

Elias PM (1983) Epidermal lipids, barrier function, and desquamation. J Invest Dermatol 80 Suppl:44s-49s

Elias PM (2005) Stratum corneum defensive functions: an integrated view. J Invest Dermatol 125:183-200

Elias PM, Feingold KR (2001) Coordinate regulation of epidermal differentiation and barrier homeostasis. Skin Pharmacol Appl Skin Physiol 14 Suppl 1:28-34

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434:857-863

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, et al. (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399-408

Feliciani C, Gupta AK, Sauder DN (1996) Keratinocytes and cytokine/growth factors. Crit Rev Oral Biol Med 7:300-318

Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1:R57-66

Findlay GH, Nurse GT, Heyl T, Hull PR, Jenkins T, Klevansky H, Morrison JG, Sher J, Schulz EJ, Swart E, Venter IJ, Whiting DA (1977) Keratolytic winter erythema or 'oudtshoorn skin': a newly recognized inherited dermatosis prevalent in South Africa. S Afr Med J 52:871-874

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806-811

Flicek P, Keibler E, Hu P, Korf I, Brent MR (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res 13:46-54

Fluckiger R, Itin PH (1993) Keratosis extremitatum (Greither's disease): clinical features, histology, ultrastructure. Dermatology 187:309-311

Francova H, Trbusek M, Zapletalova P, Kuhrova V (2004) New promoter mutations in the low-density lipoprotein receptor gene which induce familial hypercholesterolaemia phenotype: molecular and functional analysis. J Inherit Metab Dis 27:523-528

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13:1-12

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 16:949-961

Fuchs E (2007) Scratching the surface of skin development. Nature 445:834-842

Fuchs E, Raghavan S (2002) Getting under the skin of epidermal morphogenesis. Nat Rev Genet 3:199-209

Gallegos Ruiz MI, Floor K, Roepman P, Rodriguez JA, Meijer GA, Mooi WJ, Jassem E, Niklinski J, Muley T, van Zandwijk N, Smit EF, Beebe K, Neckers L, Ylstra B, Giaccone G (2008) Integration of gene dosage and gene expression in non-small cell lung cancer, identification of HSP90 as potential target. PLoS ONE 3:e0001722

Gallo RL, Nizet V (2003) Endogenous production of antimicrobial peptides in innate immunity and human disease. Curr Allergy Asthma Rep 3:402-409

Ganz T (2003) Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol 3:710-720

Gentles AJ, Karlin S (1999) Why are human G-protein-coupled receptors predominantly intronless? Trends Genet 15:47-49

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669-681

Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, Park BK, Rubenstein JL, Ekker M (2003) Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. Genome Res 13:533-543

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493-521

Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, Weber JL, Ledbetter DH, Zuffardi O (2001) Olfactory receptor-gene clusters,

genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 68:874-883

Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A, Stumm M, Tonnies H, Ventura M, Zollino M, Neri G, Barber J, Wieczorek D, Rocchi M, Zuffardi O (2002) Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. Am J Hum Genet 71:276-285

Giorda R, Ciccone R, Gimelli G, Pramparo T, Beri S, Bonaglia MC, Giglio S, Genuardi M, Argente J, Rocchi M, Zuffardi O (2007) Two classes of low-copy repeats comediate a new recurrent rearrangement consisting of duplication at 8p23.1 and triplication at 8p23.2. Hum Mutat 28:459-468

Goldstein JL, Brown MS (1990) Regulation of the mevalonate pathway. Nature 343:425-430

Gondo Y, Okada T, Matsuyama N, Saitoh Y, Yanagisawa Y, Ikeda JE (1998) Human megasatellite DNA RS447: copy-number polymorphisms and interspecies conservation. Genomics 54:39-49

Goto M, Katsumata N (2009) X-linked adrenal hypoplasia congenita caused by a novel intronic mutation of the DAX-1 gene. Horm Res 71:120-124

Gottgens B, Barton LM, Chapman MA, Sinclair AM, Knudsen B, Grafham D, Gilbert JG, Rogers J, Bentley DR, Green AR (2002) Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. Genome Res 12:749-759

Gottschaldt KM, Vahle-Hinz C (1981) Merkel cell receptors: structure and transducer function. Science 214:183-186

Groth M, Szafranski K, Taudien S, Huse K, Mueller O, Rosenstiel P, Nygren AO, Schreiber S, Birkenmeier G, Platzer M (2008) High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes. Hum Mutat 29:1247-1254

Gurnett CA, Bowcock AM, Dietz FR, Morcuende JA, Murray JC, Dobbs MB (2007) Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. Am J Med Genet A 143:27-32

Harder J, Siebert R, Zhang Y, Matthiesen P, Christophers E, Schlegelberger B, Schroder JM (1997) Mapping of the gene encoding human beta-defensin-2 (DEFB2) to chromosome region 8p22-p23.1. Genomics 46:472-475

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369-372

Hardison RC (2003) Comparative genomics. PLoS Biol 1:E58

Hollox EJ (2008) Copy number variation of beta-defensins and relevance to disease. Cytogenet Genome Res 123:148-155

Hollox EJ, Armour JA, Barber JC (2003) Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am J Hum Genet 73:591-600

Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PC, Traupe H, de Jongh G, den Heijer M, Reis A, Armour JA, Schalkwijk J (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40:23-25

Hosoi J, Murphy GF, Egan CL, Lerner EA, Grabbe S, Asahina A, Granstein RD (1993) Regulation of Langerhans cell function by nerves containing calcitonin gene-related peptide. Nature 363:159-163

Hughes JR, Cheng JF, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, de Jong P, Rubin E, Higgs DR (2005) Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proc Natl Acad Sci U S A 102:9830-9835

Hull P (1986) Keratolytic winter erythema (Oudtshoorn disease): clinical, genetic and ultrastructural aspects., University of the Witwatersrand, Johannesburg

Itin PH, Fistarol SK (2005) Palmoplantar keratodermas. Clin Dermatol 23:15-22
Iwai N, Naraba H (2005) Polymorphisms in human pre-miRNAs. Biochem Biophys Res Commun 331:1439-1444
Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318-356
Kalff-Suske M, Wild A, Topp J, Wessling M, Jacobsen EM, Bornholdt D, Engel H, Heuer H, Aalfs CM, Ausems MG, Barone R, Herzog A, Heutink P, Homfray T, Gillessen-Kaesbach G, Konig R, Kunze J, Meinecke P, Muller D, Rizzo R, Strenge S, Superti-Furga A, Grzeschik KH (1999) Point mutations throughout the GLI3 gene cause Greig cephalopolysyndactyly syndrome. Hum Mol Genet 8:1769-1777
Kalinin AE, Kajava AV, Steinert PM (2002) Epithelial barrier function: assembly and structural features of the cornified cell envelope. Bioessays 24:789-800
Kallioniemi A, Visakorpi T, Karhu R, Pinkel D, Kallioniemi OP (1996) Gene Copy Number Analysis by Fluorescence in Situ Hybridization and Comparative Genomic Hybridization. Methods 9:113-121
Kang S, Rosenberg M, Ko VD, Biesecker LG (1997) Gene structure and allelic expression assay of the human GLI3 gene. Hum Genet 101:154-157
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241-254
Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-1080
Kim S, Kim Y, Lee Y, Cho KH, Kim KH, Chung JH (2007) Cholesterol inhibits MMP-9 expression in human epidermal keratinocytes and HaCaT cells. FEBS Lett 581:3869-3874
Kim YK, Kim VN (2007) Processing of intronic microRNAs. Embo J 26:775-783
Kimyai-Asadi A, Kotcher LB, Jih MH (2002) The molecular basis of hereditary palmoplantar keratodermas. J Am Acad Dermatol 47:327-343; quiz 344-326
King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15:1051-1060
King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, Chiaromonte F, Miller W, Hardison RC (2007) Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res 17:775-786
Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76:8-32
Klopocki E, Ott CE, Benatar N, Ullmann R, Mundlos S, Lehmann K (2008) A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. J Med Genet 45:370-375
Knapp AC, Franke WW, Heid H, Hatzfeld M, Jorcano JL, Moll R (1986) Cytokeratin No. 9, an epidermal type I keratin characteristic of a special program of keratinocyte differentiation displaying body site specificity. J Cell Biol 103:657-667
Knudsen TB, Kristiansen TB, Katzenstein TL, Eugen-Olsen J (2001) Adverse effect of the CCR5 promoter -2459A allele on HIV-1 disease progression. J Med Virol 65:441-444
Koenig M, Monaco AP, Kunkel LM (1988) The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. Cell 53:219-228
Kogi M, Fukushige S, Lefevre C, Hadano S, Ikeda JE (1997) A novel tandem repeat sequence located on human chromosome 4p: isolation and characterization. Genomics 42:278-283
Komander D, Clague MJ, Urbe S (2009) Breaking the chains: structure and function of the deubiquitinases. Nat Rev Mol Cell Biol 10:550-563

Krahl D, Sigwart A, Hartschuh W, Anton-Lamprecht I, Petzoldt D (1994) [Erythrokeratolysis hiemalis. Erythematosquamous genetic dermatosis with seasonal manifestation]. Hautarzt 45:776-779

Kubo M, Hanada T, Yoshimura A (2003) Suppressors of cytokine signaling and immunity. Nat Immunol 4:1169-1176

Kunze J (1980) Neurological disorders in patients with chromosomal anomalies. Neuropediatrics 11:203-249

Langbein L, Heid HW, Moll I, Franke WW (1994) Molecular characterization of the body site-specific human epidermal cytokeratin 9: cDNA cloning, amino acid sequence, and tissue specificity of gene expression. Differentiation 55:164

Lanza G, Ferracin M, Gafa R, Veronese A, Spizzo R, Pichiorri F, Liu CG, Calin GA, Croce CM, Negrini M (2007) mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. Mol Cancer 6:54

Larsen F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. Genomics 13:1095-1107

Lee JA, Madrid RE, Sperle K, Ritterson CM, Hobson GM, Garbern J, Lupski JR, Inoue K (2006) Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. Ann Neurol 59:398-403

Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843-854

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12:1725-1735

Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A 99:7548-7553

Linzmeier RM, Ganz T (2005) Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. Genomics 86:423-430

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136-140

Lucito R, Suresh S, Walter K, Pandey A, Lakshmi B, Krasnitz A, Sebat J, Wigler M, Klein AP, Brune K, Palmisano E, Maitra A, Goggins M, Hruban RH (2007) Copy-number variants in patients with a strong family history of pancreatic cancer. Cancer Biol Ther 6:1592-1599

Lupski JR (1992) An inherited DNA rearrangement and gene dosage effect are responsible for the most common autosomal dominant peripheral neuropathy: Charcot-Marie-Tooth disease type 1A. Clin Res 40:645-652

Lupski JR (2007) Genomic rearrangements and sporadic disease. Nat Genet 39:S43-47

Lupski JR (2009) Genomic disorders ten years on. Genome Med 1:42

Manco L, Ribeiro ML, Maximo V, Almeida H, Costa A, Freitas O, Barbot J, Abade A, Tamagnini G (2000) A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. Br J Haematol 110:993-997

Margulies EH, Blanchette M, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13:2507-2518

Margulies EH, Chen CW, Green ED (2006) Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends Genet 22:187-193

Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 17:760-774

Margulies EH, Green ED (2003) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. Cold Spring Harb Symp Quant Biol 68:255-263

Martin MP, Dean M, Smith MW, Winkler C, Gerrard B, Michael NL, Lee B, Doms RW, Margolick J, Buchbinder S, Goedert JJ, O'Brien TR, Hilgartner MW, Vlahov D, O'Brien SJ, Carrington M (1998) Genetic acceleration of AIDS progression by a promoter variant of CCR5. Science 282:1907-1911

Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29-59

Matsubara H, Takeuchi T, Nishikawa E, Yanagisawa K, Hayashita Y, Ebi H, Yamada H, Suzuki M, Nagino M, Nimura Y, Osada H, Takahashi T (2007) Apoptosis induction by antisense oligonucleotides against miR-17-5p and miR-20a in lung cancers overexpressing miR-17-92. Oncogene 26:6099-6105

Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15 Spec No 1:R17-29

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374-378

McDermott DH, Zimmerman PA, Guignard F, Kleeberger CA, Leitman SF, Murphy PM (1998) CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS). Lancet 352:866-870

McLean WH, Irvine AD (2007) Disorders of keratinisation: from rare to common genetic diseases of skin and other epithelial tissues. Ulster Med J 76:72-82

Meister G, Landthaler M, Dorsett Y, Tuschl T (2004) Sequence-specific inhibition of microRNA- and siRNA-induced RNA silencing. Rna 10:544-550

Michael MZ, SM OC, van Holst Pellekaan NG, Young GP, James RJ (2003) Reduced accumulation of specific microRNAs in colorectal neoplasia. Mol Cancer Res 1:882-891

Migone TS, Humbert M, Rascle A, Sanden D, D'Andrea A, Johnston JA (2001) The deubiquitinating enzyme DUB-2 prolongs cytokine-induced signal transducers and activators of transcription activation and suppresses apoptosis following cytokine withdrawal. Blood 98:1935-1941

Moll I, Heid H, Franke WW, Moll R (1987) Distribution of a special subset of keratinocytes characterized by the expression of cytokeratin 9 in adult and fetal human epidermis of various body sites. Differentiation 33:254-265

Niemann C, Watt FM (2002) Designer skin: lineage commitment in postnatal epidermis. Trends Cell Biol 12:185-192

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302:413

Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. J Physiol 554:31-39

Okada T, Gondo Y, Goto J, Kanazawa I, Hadano S, Ikeda JE (2002) Unstable transmission of the RS447 human megasatellite tandem repetitive sequence that contains the USP17 deubiquitinating enzyme gene. Hum Genet 110:302-313

Pandit J, Danley DE, Schulte GK, Mazzalupo S, Pauly TA, Hayward CM, Hamanaka ES, Thompson JF, Harwood HJ, Jr. (2000) Crystal structure of human squalene synthase. A key enzyme in cholesterol biosynthesis. J Biol Chem 275:30610-30617

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499-502

Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100-109

Perier RC, Praz V, Junier T, Bonnard C, Bucher P (2000) The eukaryotic promoter database (EPD). Nucleic Acids Res 28:302-303

Pinkel D, Albertson DG (2005) Comparative genomic hybridization. Annu Rev Genomics Hum Genet 6:331-354

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A 99:12963-12968

Proksch E, Holleran WM, Menon GK, Elias PM, Feingold KR (1993) Barrier function regulates epidermal lipid and DNA synthesis. Br J Dermatol 128:473-482

Radhakrishna U, Wild A, Grzeschik KH, Antonarakis SE (1997) Mutation in GLI3 in postaxial polydactyly type A. Nat Genet 17:269-271

Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, Lie RT, Wilcox AJ, Fitzpatrick DR, Green ED, Mossey PA, Little J, Steegers-Theunissen RP, Pennacchio LA, Schutte BC, Murray JC (2008) Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet 40:1341-1347

Rana TM (2007) Illuminating the silence: understanding the structure and function of small RNAs. Nat Rev Mol Cell Biol 8:23-36

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, et al. (2006) Global variation in copy number in the human genome. Nature 444:444-454

Reefman E, Limburg PC, Kallenberg CG, Bijl M (2005) Apoptosis in human skin: role in pathogenesis of various diseases and relevance for therapy. Ann N Y Acad Sci 1051:52-63

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32:135-142

Reis A, Hennies HC, Langbein L, Digweed M, Mischke D, Drechsler M, Schrock E, Royer-Pokora B, Franke WW, Sperling K, et al. (1994) Keratin 9 gene mutations in epidermolytic palmoplantar keratoderma (EPPK). Nat Genet 6:174-179

Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. Nat Genet 12:288-297

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066-1073

Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2:8

Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of mammalian microRNA host genes and transcription units. Genome Res 14:1902-1910

Ruiz M, Salvador J, Bayo J, Lomas M, Moreno A, Valero M, Bernabe R, Vicente D, Jimenez J, Lopez-Ladron A (2008) Phase-II study of weekly schedule of trastuzumab, paclitaxel, and carboplatin followed by a week off every 28 days for HER2+ metastatic breast cancer. Cancer Chemother Pharmacol 62:1085-1090

Ruth K. Freinkel DW (2001) The Biology of the Skin. Parthenon Publishing Group, New York
Sakharkar MK, Chow VT, Chaturvedi I, Mathura VS, Shapshak P, Kangueane P (2004) A report on single exon genes (SEG) in eukaryotes. Front Biosci 9:3262-3267
Schwartz LM, Nambu JR, Wang Z (2002) Parkinsonism proteolysis and proteasomes. Cell Death Differ 9:479-482
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103-107
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525-528
Segre JA (2006) Epidermal barrier formation and recovery in skin disorders. J Clin Invest 116:1150-1158
Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. Annu Rev Genomics Hum Genet 7:407-442
Shimokawa O, Kurosawa K, Ida T, Harada N, Kondoh T, Miyake N, Yoshiura K, Kishino T, Ohta T, Niikawa N, Matsumoto N (2004) Molecular characterization of inv dup del(8p): analysis of five cases. Am J Med Genet A 128A:133-137
Soifer HS, Rossi JJ, Saetrom P (2007) MicroRNAs in disease and potential therapeutic applications. Mol Ther 15:2070-2079
Solis C, Aizencang GI, Astrin KH, Bishop DF, Desnick RJ (2001) Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. J Clin Invest 107:753-762
Sonkoly E, Wei T, Janson PC, Saaf A, Lundeberg L, Tengvall-Linder M, Norstedt G, Alenius H, Homey B, Scheynius A, Stahle M, Pivarcsi A (2007) MicroRNAs: novel regulators involved in the pathogenesis of Psoriasis? PLoS ONE 2:e610
Stankiewicz P, Lupski JR Structural variation in the human genome and its role in disease. Annu Rev Med 61:437-455
Starfield M, Hennies HC, Jung M, Jenkins T, Wienker T, Hull P, Spurdle A, Kuster W, Ramsay M, Reis A (1997) Localization of the gene causing keratolytic winter erythema to chromosome 8p22-p23, and evidence for a founder effect in South African Afrikaans-speakers. Am J Hum Genet 61:370-378
Strahle U, Rastegar S (2008) Conserved non-coding sequences and transcriptional regulation. Brain Res Bull 75:225-230
Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET (2007) Population genomics of human gene expression. Nat Genet 39:1217-1224
Strathdee CA, Gavish H, Shannon WR, Buchwald M (1992) Cloning of cDNAs for Fanconi's anaemia by functional complementation. Nature 356:763-767
Sutrala SR, Goossens D, Williams NM, Heyrman L, Adolfsson R, Norton N, Buckland PR, Del-Favero J (2007) Gene copy number variation in schizophrenia. Schizophr Res 96:93-99
Swensson O, Langbein L, McMillan JR, Stevens HP, Leigh IM, McLean WH, Lane EB, Eady RA (1998) Specialized keratin expression pattern in human ridged skin as an adaptation to high physical stress. Br J Dermatol 139:767-775
Takahashi K, Paladini RD, Coulombe PA (1995) Cloning and characterization of multiple human genes and cDNAs encoding highly related type II keratin 6 isoforms. J Biol Chem 270:18581-18592
Tautz D (2000) Evolution of transcriptional regulation. Curr Opin Genet Dev 10:575-579

Thienpont B, de Ravel T, Van Esch H, Van Schoubroeck D, Moerman P, Vermeesch JR, Fryns JP, Froyen G, Lacoste C, Badens C, Devriendt K (2007) Partial duplications of the ATRX gene cause the ATR-X syndrome. Eur J Hum Genet 15:1094-1097

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788-793

Tierling S, Dalbert S, Schoppenhorst S, Tsai CE, Oliger S, Ferguson-Smith AC, Paulsen M, Walter J (2006) High-resolution map and imprinting analysis of the Gtl2-Dnchc1 domain on mouse chromosome 12. Genomics 87:225-235

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23:137-144

van Deursen D, Botma GJ, Jansen H, Verhoeven AJ (2007) Comparative genomics and experimental promoter analysis reveal functional liver-specific elements in mammalian hepatic lipase genes. BMC Genomics 8:99

Vasudevan S, Tong Y, Steitz JA (2007) Switching from repression to activation: microRNAs can up-regulate translation. Science 318:1931-1934

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al. (2001) The sequence of the human genome. Science 291:1304-1351

Villard J (2004) Transcription regulation and human diseases. Swiss Med Wkly 134:571-579

Ward RD, Davis SW, Cho M, Esposito C, Lyons RH, Cheng JF, Rubin EM, Rhodes SJ, Raetzman LT, Smith TP, Camper SA (2007) Comparative genomics reveals functional transcriptional control sequences in the Prop1 gene. Mamm Genome 18:521-537

Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 102:2454-2459

Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26:225-228

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5:276-287

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562

Wilkinson KD (1997) Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. Faseb J 11:1245-1256

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:e7

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen H B, Saxild H, Nielsen C, Brunak S, Knudsen S (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology 3:research0048.1-0048.16

Yamaguchi Y, Itami S, Tarutani M, Hosokawa K, Miura H, Yoshikawa K (1999) Regulation of keratin 9 in nonpalmoplantar keratinocytes by palmoplantar fibroblasts through epithelial-mesenchymal interactions. J Invest Dermatol 112:483-488

Zhu Y, Carroll M, Papa FR, Hochstrasser M, D'Andrea AD (1996a) DUB-1, a deubiquitinating enzyme with growth-suppressing activity. Proc Natl Acad Sci U S A 93:3275-3279

Zhu Y, Lambert K, Corless C, Copeland NG, Gilbert DJ, Jenkins NA, D'Andrea AD (1997) DUB-2 is a member of a novel family of cytokine-inducible deubiquitinating enzymes. J Biol Chem 272:51-57

Zhu Y, Pless M, Inhorn R, Mathey-Prevot B, D'Andrea AD (1996b) The murine DUB-1 gene is specifically induced by the betac subunit of interleukin-3 receptor. Mol Cell Biol 16:4808-4817

Online references

Online Mendelian Inheritance in Man <http://www.ncbi.nlm.nih.gov/Omim> (Accessed January 2009)

RefSeq <http://www.ncbi.nlm.nih.gov/RefSeq/> (Accessed June 2010)

European Bioinformatics Institute <http://www.ebi.ac.uk/embl/> (Accessed May 2009)

Swiss-Prot protein knowledge database <http://www.expasy.org/sprot> (Accessed April 2009)

Protein Information Resource <http://pir.georgetown.edu> (Accessed June 2008)

Ensembl Genome Browser <http://www.ensembl.org/index.html> (Accessed August 2010)

UCSC Genome Browser <http://genome.ucsc.edu/> (Accessed August 2010)

The Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff <http://www.hgmd.cf.ac.uk/ac/> (Accessed December 2010)

The Human microRNA Disease Database <http://cmbi.bjmu.edu.cn/hmdd> (Accessed December 2010)

Basic Local Alignment Search Tool (BLAST) <http://blast.ncbi.nlm.nih.gov/Blast.cgi> (Accessed April 2010)

GeneID <http://genome.imim.es/geneid> (Accessed May 2009)

GeneScan <http://genes.mit.edu/GENSCAN> (Accessed May 2009)

Vienna RNAfold webserver <http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi> (Accessed June 2010)

MultiExperiment Viewer (MeV) <http://www.tm4.org/>

Galaxy <http://main.g2.bx.psu.edu/> (Accessed November 2009)

Appendices

Appendix A: Ethics Approval

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Ramsay

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER M070423**
PROJECT	The identification and characterisation of the causative gene for Keratolytic winter erythema in South African families
INVESTIGATORS	Prof M Ramsay
DEPARTMENT	Human Genetics
DATE CONSIDERED	07.05.04
DECISION OF THE COMMITTEE*	APPROVED UNCONDITIONALLY

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 07.05.25 **CHAIRPERSON** .....
(Professors PE Cleaton-Jones, A Dhai, M Vorster, C Feldman, A Woodiwiss)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Ramsay M Prof

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

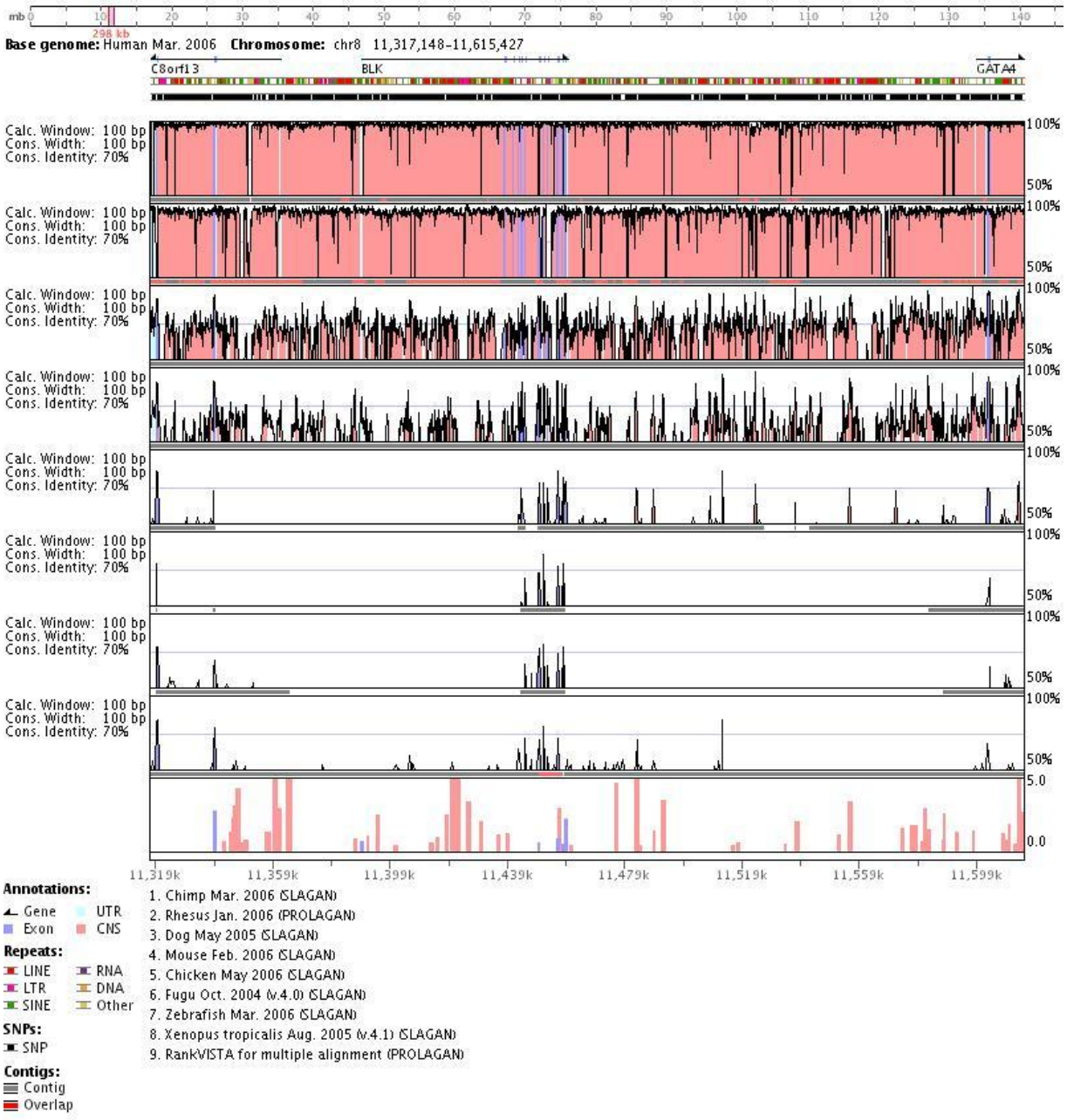
M Ramsay 10-6-07

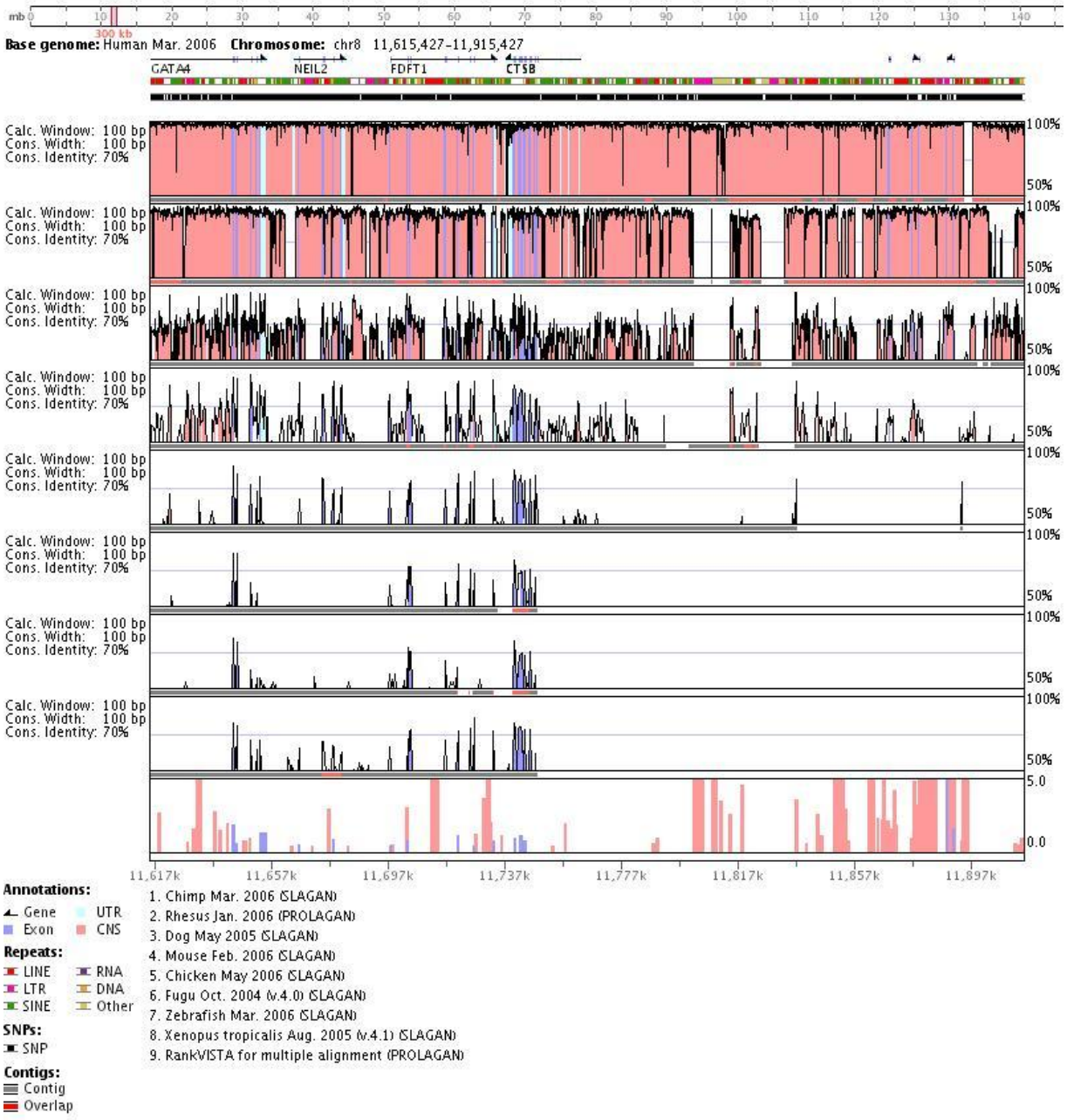
(Angela Hobbs) 04-06-07

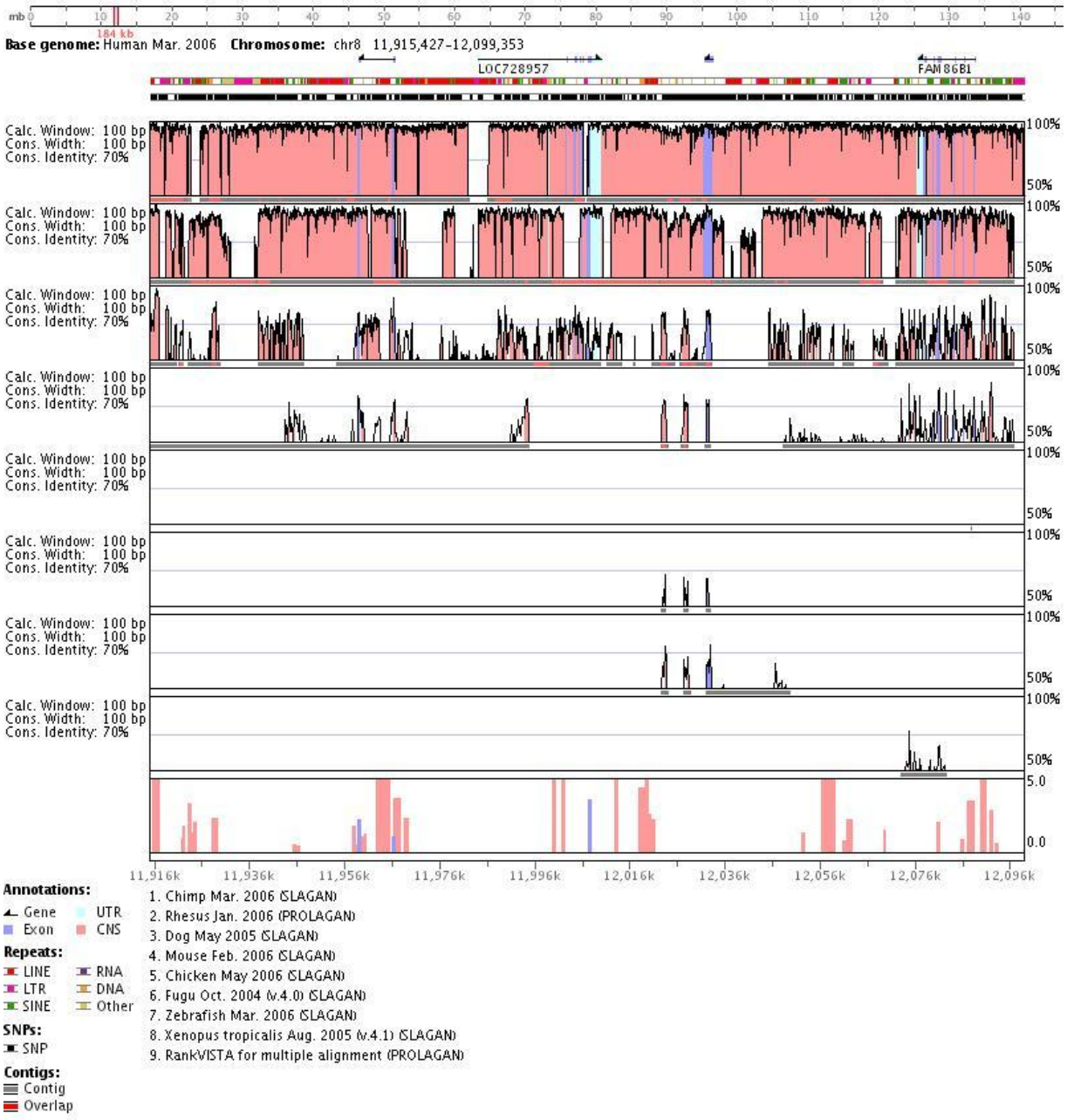
TO:0114899226 P:1

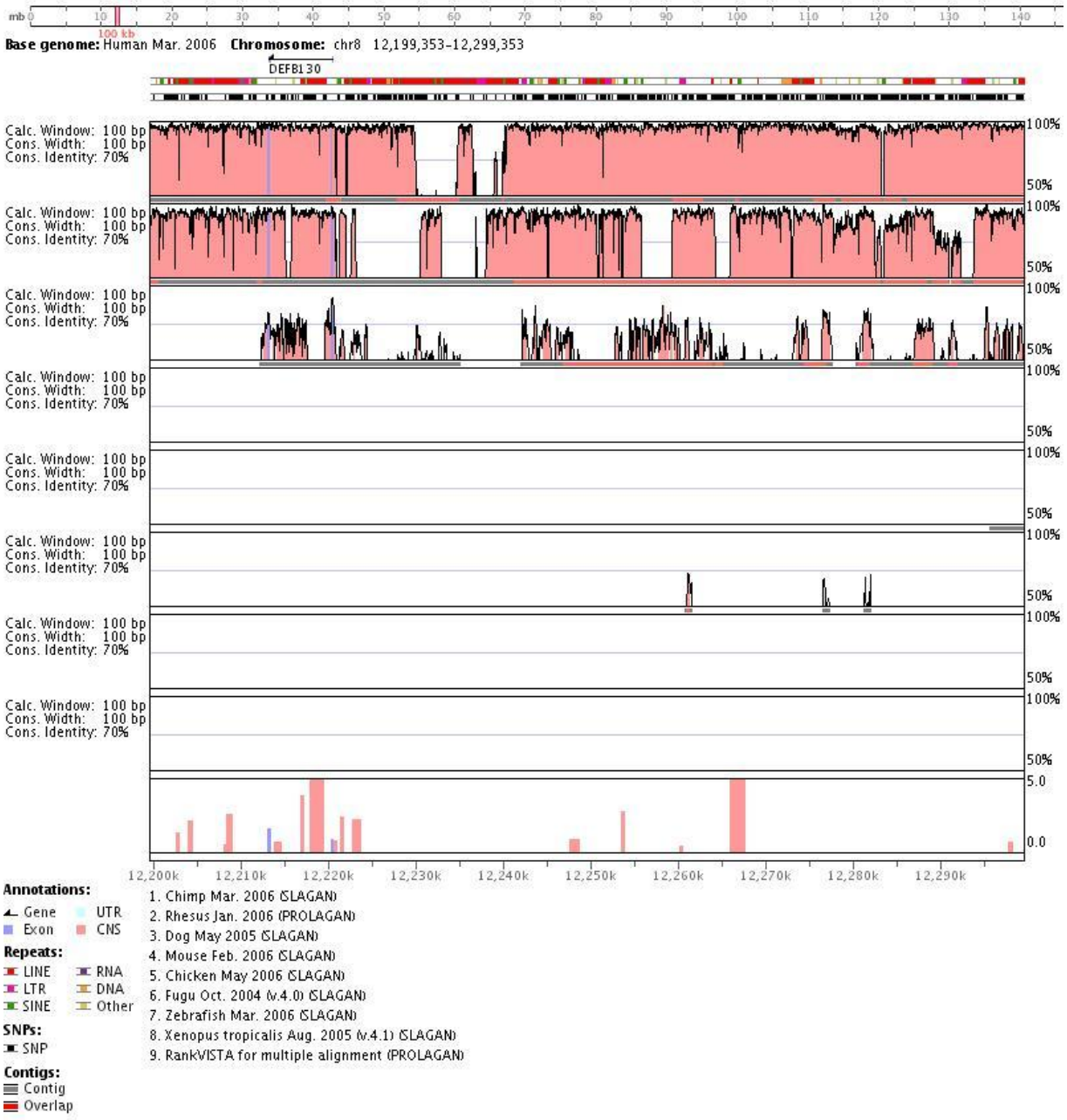
FROM: 15:02 NOV-1-2007

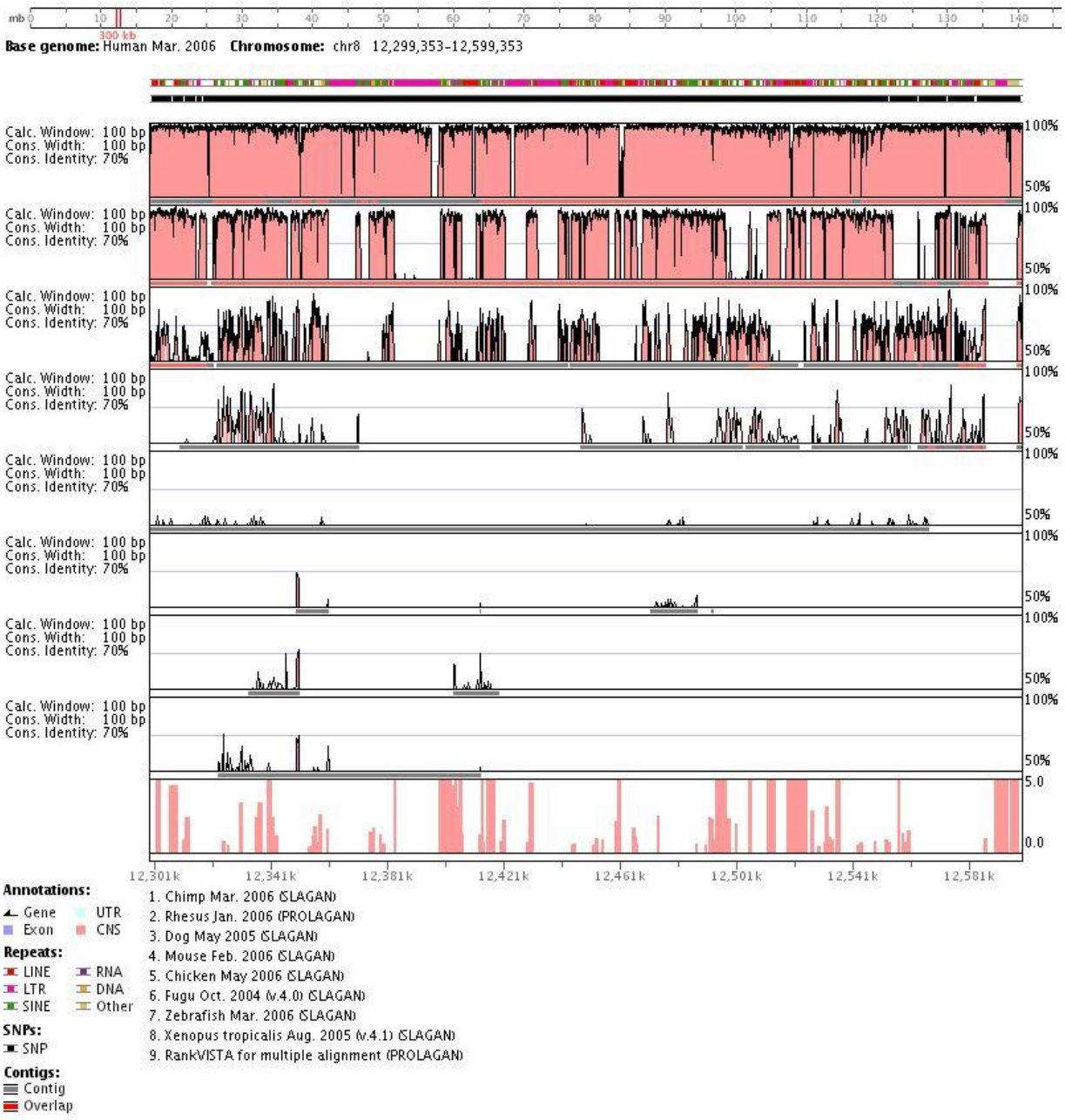
Appendix B: GenomeVISTA Plots of KWE critical region

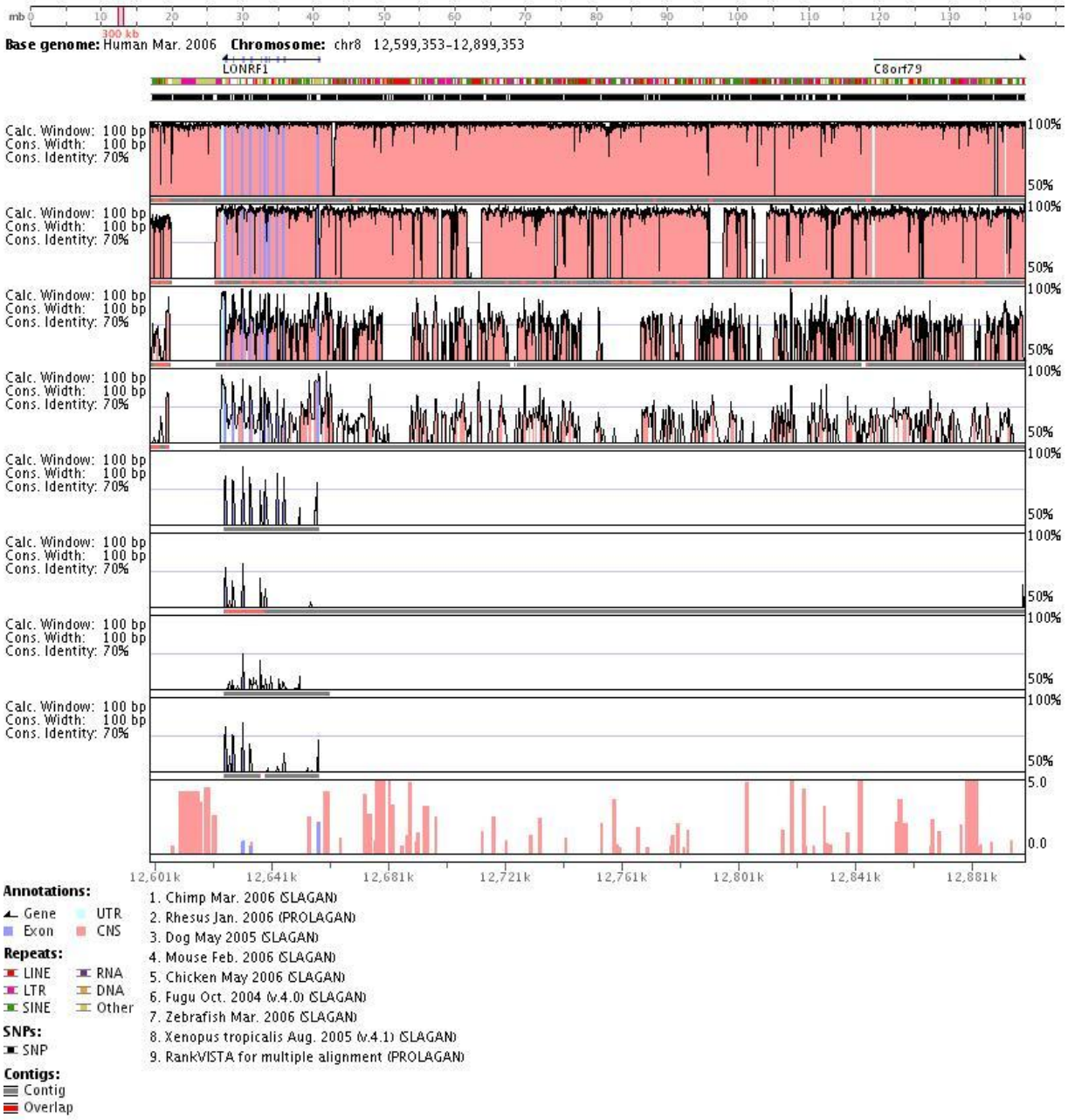



Base genome: Human Mar. 2006 Chromosome: chr8 11,615,427-11,915,427
GATA4
NEIL2
FDFT1
CTSB
Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%
100%
50%
5.0
0.0
11,617k
11,657k
11,697k
11,737k
11,777k
11,817k
11,857k
11,897k
Annotations:
Gene
Exon
UTR
CNS
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
SNPs:
SNP
Contigs:
Contig
Overlap
1. Chimp Mar. 2006 (SLAGAN)
2. Rhesus Jan. 2006 (PROLAGAN)
3. Dog May 2005 (SLAGAN)
4. Mouse Feb. 2006 (SLAGAN)
5. Chicken May 2006 (SLAGAN)
6. Fugu Oct. 2004 (v.4.0) (SLAGAN)
7. Zebrafish Mar. 2006 (SLAGAN)
8. Xenopus tropicalis Aug. 2005 (v.4.1) (SLAGAN)
9. RankVISTA for multiple alignment (PROLAGAN)


Base genome: Human Mar. 2006 Chromosome: chr8 11,915,427-12,099,353
LOC728957
FAM86B1
Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%
100%
50%
5.0
0.0
11,916k
11,936k
11,956k
11,976k
11,996k
12,016k
12,036k
12,056k
12,076k
12,096k
Annotations:
Gene
Exon
UTR
CNS
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
SNPs:
SNP
Contigs:
Contig
Overlap
1. Chimp Mar. 2006 (SLAGAN)
2. Rhesus Jan. 2006 (PROLAGAN)
3. Dog May 2005 (SLAGAN)
4. Mouse Feb. 2006 (SLAGAN)
5. Chicken May 2006 (SLAGAN)
6. Fugu Oct. 2004 (v.4.0) (SLAGAN)
7. Zebrafish Mar. 2006 (SLAGAN)
8. Xenopus tropicalis Aug. 2005 (v.4.1) (SLAGAN)
9. RankVISTA for multiple alignment (PROLAGAN)


mb 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
100 kb
Base genome: Human Mar. 2006 Chromosome: chr8 12,199,353-12,299,353
DEFB130
Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%
100%
50%
5.0
0.0
12,200k 12,210k 12,220k 12,230k 12,240k 12,250k 12,260k 12,270k 12,280k 12,290k
Annotations:
Gene
Exon
UTR
CNS
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
SNPs:
SNP
Contigs:
Contig
Overlap
1. Chimp Mar. 2006 (SLAGAN)
2. Rhesus Jan. 2006 (PROLAGAN)
3. Dog May 2005 (SLAGAN)
4. Mouse Feb. 2006 (SLAGAN)
5. Chicken May 2006 (SLAGAN)
6. Fugu Oct. 2004 (v.4.0) (SLAGAN)
7. Zebrafish Mar. 2006 (SLAGAN)
8. Xenopus tropicalis Aug. 2005 (v.4.1) (SLAGAN)
9. RankVISTA for multiple alignment (PROLAGAN)


mb 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
300 kb
Base genome: Human Mar. 2006 Chromosome: chr8 12,299,353-12,599,353
Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%
100%
50%
5.0
0.0
12,301k 12,341k 12,381k 12,421k 12,461k 12,501k 12,541k 12,581k
Annotations:
Gene
Exon
UTR
CNS
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
SNPs:
SNP
Contigs:
Contig
Overlap
1. Chimp Mar. 2006 (SLAGAN)
2. Rhesus Jan. 2006 (PROLAGAN)
3. Dog May 2005 (SLAGAN)
4. Mouse Feb. 2006 (SLAGAN)
5. Chicken May 2006 (SLAGAN)
6. Fugu Oct. 2004 (v.4.0) (SLAGAN)
7. Zebrafish Mar. 2006 (SLAGAN)
8. Xenopus tropicalis Aug. 2005 (v.4.1) (SLAGAN)
9. RankVISTA for multiple alignment (PROLAGAN)


Base genome: Human Mar. 2006 Chromosome: chr8 12,599,353-12,899,353
LONRF1
C8orf79
Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%
100%
50%
5.0
0.0
12,601k
12,641k
12,681k
12,721k
12,761k
12,801k
12,841k
12,881k
Annotations:
Gene
Exon
UTR
CNS
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
SNPs:
SNP
Contigs:
Contig
Overlap
1. Chimp Mar. 2006 (SLAGAN)
2. Rhesus Jan. 2006 (PROLAGAN)
3. Dog May 2005 (SLAGAN)
4. Mouse Feb. 2006 (SLAGAN)
5. Chicken May 2006 (SLAGAN)
6. Fugu Oct. 2004 (v.4.0) (SLAGAN)
7. Zebrafish Mar. 2006 (SLAGAN)
8. Xenopus tropicalis Aug. 2005 (v.4.1) (SLAGAN)
9. RankVISTA for multiple alignment (PROLAGAN)

Appendix C: Coordinates of orthologous regions for each of the highly conserved sequences

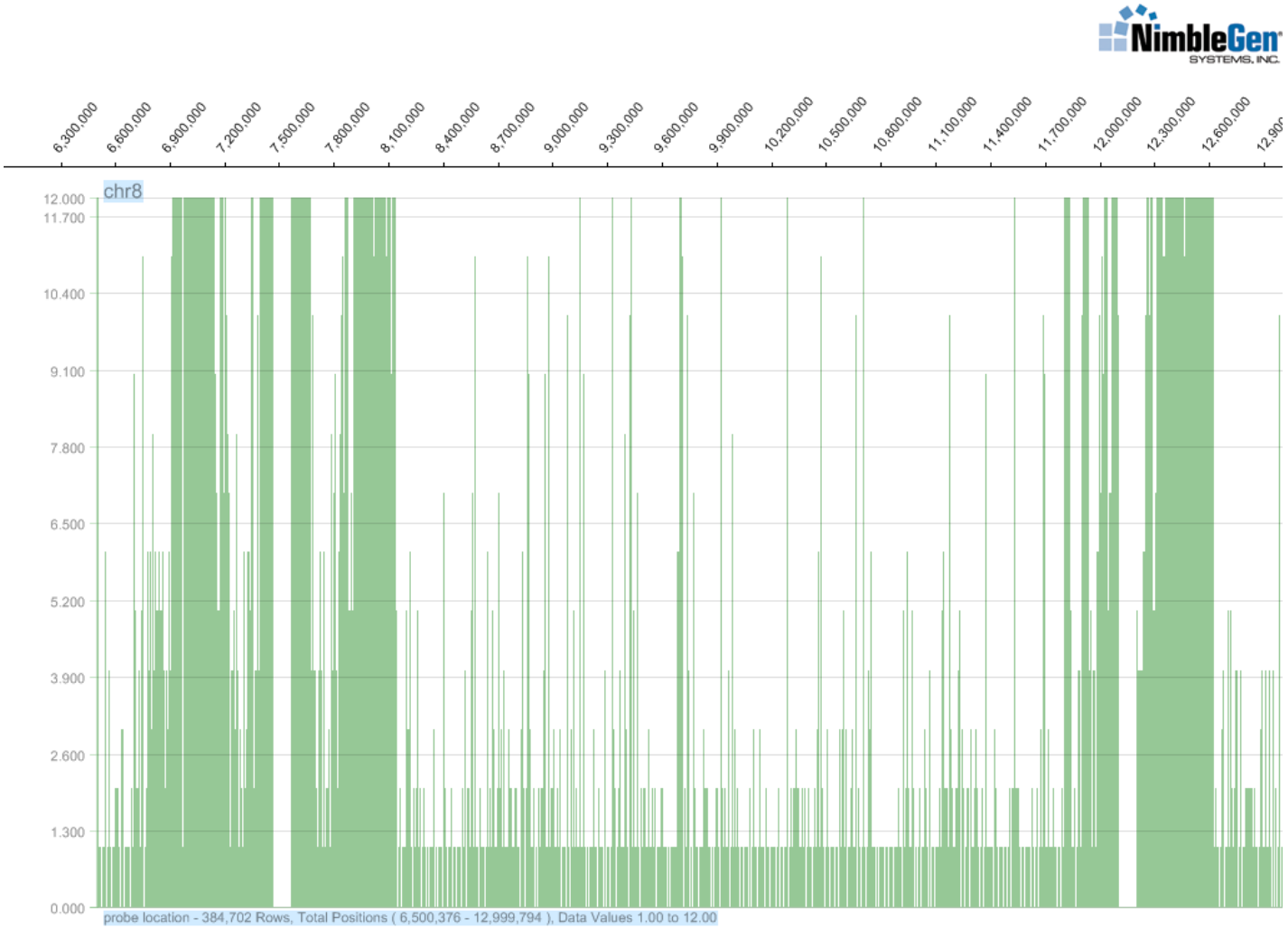
Organism	Human Chromosomal Coordinates		Ortholgous Coordinates		Sequence Length (Bp)	% Conservation	Type
	Start	End	Start	End			
C8orf13 - Region 1							
Chimp	Chr8: 11 319 292	Chr8: 11 319 555	Contig 187.40:886	Contig 187.40:1149	264	98.50%	Exon
Dog	Chr8: 11 319 292	Chr8: 11 319 553	Chr25:29 467 282	Chr25:29 467 021	262	88.20%	Exon
Mouse	Chr8: 11 319 292	Chr8: 11 319 552	Chr14: 62 416 750	Chr14: 62 416 490	261	88.50%	Exon
Chicken	Chr8: 11 319 292	Chr8: 11 319 552	Chr3: 110 013 246	Chr:3 110 013 506	261	84.70%	Exon
Fugu	Chr8: 11 319 242	Chr8: 11 319 552	Scaffold_72 : 128 526	Scaffold_72: 128 812	259	74.50%	Exon
Zebrafish	Chr8: 11 319 292	Chr8: 11 319 552	Zv6_Scaffold3 733: 552	Zv6_Scaffold37 33: 295	262	71.00%	Exon
	Chr8: 11 319 294	Chr8: 11 319 552	Chr20 : 20 052 962	Chr20: 20 052 703	260	70.00%	Exon
Frog	Chr8: 11 319 302	Chr8: 11 319 555	Scaffold_63: 2 788 622	Scaffold_63: 2 788 369	254	84.30%	Exon
C8orf13 - Region 2							
Chimp	Chr8: 11 338 950	Chr8: 11 339 330	Contig 187.36: 13 100	Contig 187.36:13 480	381	99.50%	Exon
Rhesus	Chr8: 11 338 950	Chr8: 11 339 330	Chr8: 9 306 543	Chr8: 9 306 163	381	96.90%	Exon
Dog	Chr8: 11 338 950	Cgr8: 11 339 330	Chr25: 29 445 407	Chr25: 29 445 039	387	78.30%	Exon
Mouse	Chr8: 11 338 950	Chr8: 11 339 330	Chr14: 62 406749	Chr14: 62 406 6366	384	79.90%	Exon
Chicken	Chr8: 11 339 134	Chr8: 11 339 291	Chr3: 110 023 050	Chr3: 110 023 207	158	70.30%	Exon
Fugu	Chr8: 11 339 184	Chr8: 11 339 261	Scaffold_72: 127 884	Scaffold_72: 127 155	272	82.10%	Exon
Frog	Chr8: 11 339 138	Chr8: 11 339 221	Scaffold_63: 2 782 865	Scaffold_63: 2 782 712	154	66.90%	Exon

Region 3							
Chimp	Chr8: 11 512 454	Chr8: 11 512 859	Contig 187.22: 3998	Contig 187.22: 4403	406	98.30%	Inter genic
Rhesus	Chr8: 11 512 454	Chr8: 11 512 859	Chr8: 9 126 851	Chr8: 9 126 438	414	93.20%	Inter genic
Dog	Chr8: 11 512 465	Chr8: 11 512 576	Chr25: 29 271 104	Chr25: 29 270 993	112	82.10%	Inter genic
Mouse	Chr8: 11 512 468	Chr8: 11 512 814	Chr14: 62 293 119	Chr14: 62 292 779	350	84.90%	Inter genic
Chicken	Chr8: 11 512 555	Chr8: 11 512 835	Chr3: 110 100 878	Chr3: 110 101 160	287	76.00%	Inter genic
Frog	Chr8: 11 512 646	Chr8: 11 512 773	Scaffold_63: 2 659 858	Scaffold_63: 2 659 727	132	82.60%	Inter genic
Region 4							
Chimp	Chr8: 12 023 858	Chr8: 12 024 011	Chr12: 8 524 857	Chr12: 8 525 010	154	90.30%	Inter genic
Rhesus	Chr8: 12 023 858	Chr8: 12 024 011	Chr8: 8 342 521	Chr8: 8 342 368	154	90.90%	Inter genic
Dog	Chr8: 12 023 886	Chr8: 12 024 011	ChrUn: 70 218 979	ChrUn: 70 219 104	126	75.40%	Inter genic
	Chr8: 12 023 886	Chr8: 12 024 033	Chr16: 57 021 906	Chr16: 57 022 050	148	75.00%	Inter genic
Mouse	Chr8: 12 023 824	Chr8: 12 023 369	Chr7: 103 125 240	Chr7: 103 125 385	146	71.90%	Inter genic
	Chr8: 12 023 824	Chr8: 12 023 969	Chr7: 103 291 625	Chr7: 103 291 480	146	71.90%	Inter genic
Fugu	Chr8: 12 023 887	Chr8: 12 023 985	Scaffold_115: 724 663	Scaffold_115:7 24 565	99	71.70%	Inter genic
Zebrafish	Chr8: 12 023 859	Chr8: 12 023 981	Chr12: 15 213 452	Chr12: 15 213 574	123	74.80%	Inter genic

Region 5							
Chimp	Chr8: 12 028 566	Chr8: 12 028 689	Chr12: 8 520 137	Chr12: 8 520 260	124	91.90%	Inter genic
Rhesus	Chr8: 12 028 566	Chr8: 12 028 689	Chr8: 8 342 525	Chr8: 8 342 402	124	85.50%	Inter genic
Dog	Chr8: 12 028 577	Chr8: 12 028 687	Chr16: 57 066 600	Chr16: 57 066 710	111	72.10%	Inter genic
Mouse	Chr8: 12 028 574	Chr8: 12 028 687	Chr7: 103 125 278	Chr7: 103 125 391	114	74.60%	Inter genic
Zebrafish	Chr8: 12 028 577	Chr8: 12 028 690	Chr12: 15 213 458	Chr12: 15 213 571	114	71.10%	Inter genic
Region 6							
Chimp	Chr8: 12 032 672	Chr8: 12 032 771	Contig82435.1: 170	Contig82435.1: 71	100	98.00%	Inter genic
	Chr8: 12 032 672	Chr8: 12 032 771	Chr12: 8 524 209	Chr12: 8 524 308	100	98.00%	Inter genic
Rhesus	Chr8: 12 032 674	Chr8: 12 032 771	Chr8: 8 343 168	Chr8: 8 343 070	98	98.00%	Inter genic
Dog	Chr8: 12 032 672	Chr8: 12 032 771	Chr16: 57 065 941	Chr16: 57 066 040	100	81.00%	Inter genic
	Chr8: 12 032 672	Chr8: 12 032 771	ChrUn: 70 228 241	ChrUn: 70 228 247	100	77.00%	Inter genic
Mouse	Chr8: 12 032 672	Chr8: 12 032 771	Chr:7 103 292 246	Chr7: 103 292 147	100	74.00%	Inter genic
Fugu	Chr8: 12 032 670	Chr8; 12 032 766	Scaffold_115: 727 739	Scaffold_115: 727 643	100	70.00%	Inter genic
Region 7							
Chimp	Chr8: 12 033 308	Chr8: 12 033 490	Chr12: 8 524 844	Chr12: 8 525 026	183	94.50%	Inter genic
Rhesus	Chr8: 12 033 308	Chr8: 12 033 489	Chr8: 8 342 534	Chr8: 8 342 353	182	91.20%	Inter genic

Dog	Chr8: 12 033 308	Chr8: 12 033 490	Chr16: 57 021 865	Chr16: 57 022 044	183	72.10%	Inter genic
Mouse	Chr8: 12 033 311	Chr8: 12 033 438	Chr7: 103 291 601	Chr7: 103 291 474	128	73.40%	Inter genic
Fugu	Chr8: 12 033 340	Chr8: 12 033 448	Scaffold_115: 724 673	Scaffold_115: 724 565	109	71.60%	Inter genic
Region 8							
Chimp	Chr8: 12 412 983	Chr8: 12 413 244	Chr8: 7 739 492	Chr8: 7 739 231	262	94.30%	Inter genic
Dog	Chr8: 12 412 939	Chr8: 12 413 209	Chr6: 39 360 063	Chr6: 39 359 796	274	74.80%	Inter genic
Fugu	Chr8:12 412 997	Chr8: 12 413 163	Scaffold_450: 91 639	Scaffold_450:9 1474	167	73.70%	Inter genic
Zebrafish	Chr8: 12412 996	Chr8: 12 413 187	Chr20: 27 726 082	Chr20: 27 726 275	194	69.60%	Inter genic

Appendix D: Probe coverage of the custom Nimblegen-Roche CGH tiling array



Appendix E: Summary of the pairwise copy number variation analysis

CNV regions present only in affected samples in 100bp averaged dataset based on pairwise comparisons									
Duplicated Regions									
Start	**Stop**	**Start**	**Stop**	**Start**	**Stop**	**Start**	**Stop**		
8975064	8975085	7133722	7133867	7419057	7419144	7660772	7661468		
9054010	9054036	7142566	7142601	7419444	7419476	7661468	7661612		
9359360	9359388	7145064	7145085	7419476	7419826	7662093	7662116		
11773425	11773452	7145793	7145816	7419826	7420056	7662116	7662167		
12657336	12657360	7217733	7217760	7420079	7420556	7662753	7662801		
		7288762	7288787	7421576	7421807	7665852	7665876		
Deleted Regions		7373939	7373960	7425444	7425476	7666352	7666376		
Start	**Stop**	7394553	7394592	7425476	7425824	7668360	7668407		
6575099	6575196	7394708	7394733	7426173	7426197	7792520	7792548		
6826232	6826260	7395228	7395252	7426511	7426583	7813392	7813415		
6844378	6844401	7395417	7395453	7426703	7426796	7816127	7816151		
6845016	6845052	7395896	7396040	7429212	7429620	7904372	7904397		
6863480	6863508	7396532	7396620	7607879	7607964	7907601	7907662		
6864178	6864203	7398708	7398824	7610148	7610364	7908514	7908692		
7101356	7101622	7398849	7398956	7610372	7610783	7909065	7909164		
7102906	7103000	7399368	7399389	7611441	7611728	7909788	7909812		
7103204	7103232	7399389	7399496	7611742	7611766	7910072	7910244		
7103544	7103904	7402294	7402316	7611766	7611909	7910457	7910674		
7110610	7110803	7403568	7403700	7619412	7619543	7912713	7912762		
7110812	7110992	7406360	7406591	7620320	7620416	7915188	7915488		
7111161	7111548	7407036	7407191	7623540	7623864	7915497	7915595		
7116850	7117068	7407200	7407240	7627964	7628062	7915604	7915668		
7118252	7118424	7409891	7409951	7634385	7634663	7915668	7915689		

	7118444	7118614		7414724	7414784		7638524	7638670		7915689	7915820		
	7124460	7124480		7417220	7417426		7639164	7639416		7916517	7916552		
	7125777	7125861		7417511	7417572		7648235	7648416		7916552	7916649		
	7126066	7126222		7417655	7417676		7650908	7651005		7917560	7917887		
	7132092	7132316		7417784	7418171		7653119	7653718		7918103	7918160		
	7133337	7133374		7418340	7418361		7654464	7654714		7919900	7919952		
	7133674	7133709		7418640	7418818		7658554	7658660		7920368	7920432		

	Start	Stop		Start	Stop		Start	Stop			
	7923227	7923252		10078016	10078044		11857856	11857881			
	7923297	7923321		10318776	10318857		11972073	11972098			
	7923704	7923804		10345210	10345234		12069203	12069249			
	7924283	7924304		10368492	10368516		12070416	12070439			
	7925088	7925111		10413189	10413215		12318620	12318644			
	7927968	7928312		10433289	10433312		12844882	12844904			
	8239772	8239800		10562516	10562580		12969776	12969801			
	8358600	8358624		10571493	10571517						
	8426300	8426326		10618880	10618905						
	8442332	8442360		10740597	10740621						
	8599223	8599248		10849700	10849724						
	8652120	8652141		10854693	10854716						
	8871656	8871683		10899000	10899020						
	8891746	8891769		10917560	10917620						
	8945528	8945556		11002668	11002690						
	9014180	9014204		11096624	11096648						
	9022822	9022844		11200244	11200272						
	9050015	9050036		11207492	11207553						
	9089927	9089952		11242308	11242332						
	9135093	9135116		11243484	11243505						
	9373560	9373584		11251845	11251870						
	9466892	9466919		11355104	11355128						
	9655821	9655848		11359604	11359631						
	9692277	9692303		11426088	11426112						
	9815460	9815483		11519844	11519865						
	9883124	9883152		11655492	11655512						
	9903668	9903696		11731317	11731342						
	10013637	10013662		11841560	11841585						

Appendix F: CNV Pairwise Comparison Results

Red blocks indicate CNV regions identified in more than one of the affected\unaffected pairs

100bp Averaged Pairwise Comparisons Duplicated Regions												
9302_9102		9402_8902		8002_7902		5202_7602		2502_2402				
Start	Stop	Start	Stop	Start	Stop	Start	Stop	Start	Stop			
6585300	6585400	6579300	6579400	8294900	8295300	7988100	7988300	6579300	6579400			
6936500	6936600	6922900	6923200	9372000	9372100	8615400	8615500	6976600	6976700			
7003700	7003800	6958900	6959000	9550700	9553200	11188500	11188600	6978500	6978600			
7067300	7067400	7020450	7019400	11435900	11436200	11770400	11770700	7022800	7022900			
7067600	7067700	7026400	7026500	11949800	11949900			7045300	7045400			
7112200	7112500	7045400	7045500	11952200	11952300			7179200	7179300			
7177800	7178000	7204900	7205000	12228300	12228700			7204800	7205100			
7179900	7180000	7441000	7441100	12228800	12229000			7381600	7381700			
7182500	7182700	7984200	7984300					7845200	7845300			
7582000	7582100	7992500	7992600					7981900	7982000			
7611300	7611400	8170800	8170900					7997900	7998600			
7628400	7628800	8975000	8975100					7999600	7999700			
7633800	7634100	9080000	9080300					8170800	8170900			
7643700	7644100	9450200	9450300					8975000	8975100			
7656800	7657000	9703700	9703800					9550700	9553200			
7862400	7862500	9769400	9769700					10045400	10045500			
7863300	7863400	10232900	10233000					11082700	11082800			
7867400	7867500	10429500	10429600					11283200	11283300			
7867600	7867700	10478900	10479100					11773200	11773500			
7872300	7872400	10719100	10719200					11824800	11824900			
7922900	7923200	11034700	11034900					11965000	11965200			

8133800	8134100	11059500	11059600					12005900	12006000			
8615400	8615500	11082700	11082800					12249800	12250200			
11268900	11269200	11253900	11254000					12259100	12259200			
11338300	11338400	11269900	11270200					12278300	12278400			
11344900	11345000	11770700	11771700					12475600	12475900			
12024000	12024100	11824800	11824900					12514300	12514400			
12027200	12027300	11929700	11930100					12540000	12540100			
12027900	12028000	12208100	12208200					12602200	12602300			
12031000	12031100	12218500	12218700					12775700	12775800			
12032800	12032900	12460100	12460200									
12033100	12033200	12481900	12482000									
12033400	12033500	12514200	12514300									
12276500	12276600	12540000	12540100									
12277000	12277200	12582900	12583000									
12281500	12281600	12587300	12587400									
12283800	12283900	12775700	12775800									
12819700	12820000											
12866500	12866700											

100bp Averaged Pairwise Comparisons Deleted Regions												
9302_9102		9402_8902		8002_7902		5202_7602		2502_2402				
Start	Stop	Start	Stop	Start	Stop	Start	Stop	Start	Stop			
6551700	6551800	6700000	6700200	7068200	7068300	7794100	7794200	6575100	6575200			
6697300	6697400	6855500	6855500	7072600	7072700	9184400	9185200	6845200	6845300			
6700000	6700100	6915000	6915000	7103300	7103900	9403500	9403600	6846900	6847000			
6820400	6820600	6932900	6932900	7106300	7106600	10728500	10728600	7099100	7099400			
6839500	6839700	7019300	7019300	7116000	7116100	11506200	11506400	7100700	7100800			
6858600	6858700	7077300	7077500	7129500	7129700	11640200	11640500	7101400	7102200			
6863400	6863500	7099700	7100000	7131200	7131300	11722800	11723200	7102600	7102800			
6864100	6864200	7100100	7100600	7134100	7134500			7102900	7103500			
7238700	7238900	7100900	7101000	7141800	7141900			7103600	7104000			
7756000	7756100	7101200	7101800	7144700	7144800			7104100	7104500			
7756200	7756300	7101900	7102600	7145100	7145200			7104900	7105200			
7816100	7816200	7102700	7102900	7262000	7262100			7109300	7109800			
8543100	8543200	7104000	7104100	7262300	7262400			7110000	7110500			
9403500	9403600	7104300	7105400	7286700	7286900			7110600	7110700			
9686500	9686800	7106300	7106600	7287100	7287200			7110800	7111700			
10561400	10561600	7107200	7107600	7289600	7289900			7111900	7112500			
12396100	12396300	7107800	7108200	7290000	7290100			7115000	7115500			
12712900	12713200	7108300	7108400	7307200	7307300			7116600	7117500			
		7108500	7108900	7307600	7307700			7117600	7118000			
		7109900	7110200	7307800	7307900			7118200	7118600			
		7110500	7110600	7313500	7313600			7118900	7119800			
		7111200	7111700	7330500	7330600			7124300	7125100			
		7111800	7112900	7331100	7331200			7125200	7126400			
		7113700	7114100	7354200	7354300			7126500	7127300			
		7114900	7115300	7371100	7371150			7127500	7127900			
		7115400	7115800	7371300	7371400			7130600	7130800			

		7116000	7116100	7371500	7371700			7131200	7131300			
		7116200	7116300	7371900	7372300			7131900	7132100			
		7116400	7117000	7372500	7372600			7132200	7132700			
		7117100	7117900	7373000	7373100			7132800	7133300			
		7118000	7118200	7385800	7385900			7133400	7133700			
		7118700	7118800	7386200	7386300			7133800	7134000			
		7118900	7119300	7386500	7386900			7134100	7134500			
		7119500	7120600	7388500	7388600			7134600	7135000			
		7121200	7121900	7395800	7395900			7137500	7137600			
		7122500	7122900	7398900	7399000			7139500	7140300			
		7123000	7123200	7403600	7403700			7140700	7140900			
		7123300	7123500	7405600	7405700			7141100	7141200			
		7123800	7123900	7410200	7410300			7141300	7141400			
		7124000	7124600	7410400	7410500			7142300	7143000			
		7124800	7125500	7413700	7414000			7144800	7144900			
		7125600	7125800	7417500	7417600			7391700	7391800			
		7126800	7126900	7421600	7421800			7394200	7394900			
		7127400	7128300	7429300	7429600			7395000	7395100			
		7128400	7128500	7455100	7455200			7395600	7395700			
		7128900	7129500	7576800	7576900			7395900	7396100			
		7130100	7130200	7582700	7582800			7396700	7397400			
		7130700	7131100	7582900	7583100			7398600	7398900			
		7131200	7131300	7608100	7608200			7399400	7399500			
		7131400	7131500	7611300	7611400			7401800	7402600			
		7131700	7132000	7616900	7617000			7402800	7402900			
		7132800	7133100	7620500	7620600			7403100	7403200			
		7133200	7133400	7626100	7626300			7403300	7403700			
		7133800	7134500	7635800	7636000			7404300	7405100			
		7134800	7135300	7639200	7639400			7405600	7405700			

		7135400	7135900	7640600	7641300			7406400	7406500			
		7136600	7136700	7643100	7643200			7407100	7407200			
		7136800	7137100	7655800	7656000			7409500	7410000			
		7137600	7138100	7656700	7656900			7410900	7411100			
		7138300	7138500	7658000	7658400			7411200	7411400			
		7138600	7138900	7665000	7665200			7412000	7412100			
		7139000	7139100	7671000	7671100			7414000	7414300			
		7139200	7139900	7672300	7672400			7416700	7417200			
		7140400	7140600	7673200	7673300			7417500	7417700			
		7140700	7140900	7674000	7674100			7417800	7418100			
		7141000	7141100	7674200	7674300			7418300	7418800			
		7141500	7141600	7674400	7674600			7418900	7419100			
		7141700	7141800	7687600	7687700			7419500	7420300			
		7142000	7142500	7688500	7688800			7421600	7421800			
		7142600	7143000	7688900	7689000			7424400	7424700			
		7199650	7143400	7689100	7689300			7424800	7425800			
		7144700	7144800	7689400	7689500			7426000	7426200			
		7145000	7145100	7720200	7720300			7426300	7426800			
		7145300	7145600	7736700	7736800			7427200	7428200			
		7145800	7145900	7743000	7743100			7429200	7429300			
		7238100	7238200	7743500	7743600			7607600	7608000			
		7246000	7246100	7743700	7743800			7608300	7608600			
		7261700	7261800	7744000	7744200			7609900	7610700			
		7261900	7262000	7746500	7746600			7611200	7611400			
		7269200	7269300	7759200	7759300			7611500	7612000			
		7272400	7272500	7760800	7760900			7612200	7613400			
		7287000	7287100	7761300	7761400			7617700	7618400			
		7288700	7288900	7761700	7762000			7618500	7618800			
		7290700	7291000	7764600	7764700			7618900	7619000			

		7304600	7304700	7789400	7789500			7619300	7619700			
		7371200	7371200	7804900	7805000			7619800	7620200			
		7373400	7373500	7805200	7805300			7620300	7620700			
		7374100	7374200	7810100	7810200			7623600	7624100			
		7385800	7385900	7812500	7812600			7625300	7626000			
		7387600	7387800	7816600	7816700			7626200	7626600			
		7391300	7391700	7905000	7905200			7626900	7627200			
		7392000	7392300	7925200	7926000			7627400	7627800			
		7392500	7392600	7952900	7953000			7627900	7628300			
		7392800	7392900	8110100	8110600			7630800	7630900			
		7393300	7393500	8942400	8942600			7633000	7634100			
		7393600	7394100	9709400	9709900			7634200	7634300			
		7394200	7394500	10654400	10654500			7634400	7634800			
		7394700	7394800	12064900	12065000			7634900	7636000			
		7395100	7395200	12065500	12065600			7640600	7641300			
		7395300	7395600	12068000	12068100			7641500	7641700			
		7395800	7396000	12069700	12069800			7641800	7641900			
		7396300	7396600	12072300	12072400			7642000	7642100			
		7396900	7397900	12314100	12314300			7642300	7642600			
		7398000	7398300	12314800	12314900			7642700	7643300			
		7398400	7398600	12317900	12318000			7645500	7646100			
		7399100	7399300	12504600	12504700			7646800	7647000			
		7399800	7400500	12510700	12510800			7648200	7649000			
		7400800	7400900					7649500	7649600			
		7401100	7401700					7649900	7650000			
		7401800	7401900					7650400	7650700			
		7402000	7402200					7650800	7651300			
		7402300	7402400					7654500	7654700			
		7402900	7403200					7655800	7656600			

		7403400	7403700					7657100	7657300			
		7403900	7404200					7657500	7657700			
		7404600	7405500					7658400	7659200			
		7405600	7406000					7660800	7661400			
		7406100	7406300					7662100	7662200			
		7406700	7406800					7664400	7664600			
		7406900	7407000					7664800	7664900			
		7407900	7408200					7665200	7665400			
		7408500	7408600					7665900	7666000			
		7408700	7409400					7666100	7666200			
		7409500	7409600					7666300	7666500			
		7410000	7410300					7666700	7667100			
		7410400	7410500					7669000	7669200			
		7410600	7410900					7686700	7686800			
		7411100	7411200					7904400	7904500			
		7411300	7411400					7907200	7907800			
		7411600	7411900					7907900	7908500			
		7412200	7413200					7908600	7909200			
		7413300	7413600					7909300	7909500			
		7414200	7414600					7909900	7910600			
		7414800	7415300					7912700	7913200			
		7415400	7415800					7915100	7915800			
		7415900	7416100					7916400	7917100			
		7416500	7417200					7917500	7918200			
		7417300	7417500					7919700	7919800			
		7417600	7417800					7920300	7920400			
		7418400	7418500					7922400	7923700			
		7418700	7419100					7924000	7924600			
		7419200	7419800					7924700	7925000			

		7420100	7420700					7925200	7926000			
		7420900	7421500					7928000	7928300			
		7422000	7422100					8884700	8885000			
		7422200	7422300					9342200	9342300			
		7422500	7423400					9372000	9372100			
		7423500	7423600					9719800	9719900			
		7423800	7424000					10851400	10851500			
		7424100	7425200					11045500	11045700			
		7425400	7425800					11352100	11352200			
		7425900	7426000					11446800	11446900			
		7426100	7426200					12067600	12067700			
		7426400	7426800					12067900	12068000			
		7426900	7427200					12068100	12068200			
		7427500	7428700					12068300	12068500			
		7428800	7429100					12069300	12069400			
		7429700	7429900					12069800	12070000			
		7442600	7442600					12070400	12070900			
		7607300	7607500					12316800	12317000			
		7608000	7608100					12317100	12317800			
		7608200	7608500					12317900	12318000			
		7608700	7609200					12318600	12318700			
		7609300	7609600					12318900	12319200			
		7609700	7610300					12319700	12320000			
		7610400	7611100					12320100	12320200			
		7611200	7611300					12353300	12353500			
		7611800	7612100					12742100	12742400			
		7612500	7612600									
		7612800	7614000									
		7614500	7615100									

		7615800	7616200									
		7616300	7616800									
		7616900	7617000									
		7617100	7617500									
		7618500	7618800									
		7618900	7619300									
		7619500	7619700									
		7620700	7621800									
		7622100	7622200									
		7623300	7623800									
		7623900	7624500									
		7624700	7625200									
		7626200	7626400									
		7626500	7626800									
		7627100	7627300									
		7628100	7629600									
		7629800	7630100									
		7630900	7631000									
		7631100	7631500									
		7631600	7632300									
		7632400	7632900									
		7633800	7634100									
		7634200	7634600									
		7634800	7635000									
		7635100	7635500									
		7636000	7637100									
		7637400	7637500									
		7637700	7637800									
		7638700	7639100									

		7639300	7639800									
		7640000	7640700									
		7641500	7641700									
		7641800	7642200									
		7642400	7642600									
		7643400	7644400									
		7644500	7644800									
		7645100	7645200									
		7645300	7646000									
		7646200	7646300									
		7646400	7646700									
		7646900	7647600									
		7647700	7648200									
		7649100	7649400									
		7649500	7649700									
		7651000	7651200									
		7651300	7652400									
		7652700	7652800									
		7653000	7653100									
		7654100	7654400									
		7654600	7655100									
		7655300	7655700									
		7656800	7657000									
		7657100	7657400									
		7657700	7657900									
		7658400	7658500									
		7658700	7659200									
		7659300	7660000									
		7660400	7660600									

		7661700	7662000									
		7662300	7662900									
		7663000	7664000									
		7664400	7664700									
		7664800	7665100									
		7665400	7665500									
		7665600	7665700									
		7666100	7666200									
		7666300	7666500									
		7666600	7666700									
		7666800	7667600									
		7668000	7668200									
		7668400	7668500									
		7668600	7668800									
		7669300	7669500									
		7686350	7674900									
		7715100	7715200									
		7762600	7762700									
		7763600	7763700									
		7789000	7789100									
		7810100	7810200									
		7904600	7904800									
		7904900	7905200									
		7905400	7905500									
		7907200	7907600									
		7907700	7907900									
		7908600	7908700									
		7909100	7909300									
		7909400	7909500									

		7909600	7910000									
		7911100	7911200									
		7911300	7911500									
		7911600	7912000									
		7912100	7912600									
		7912800	7913900									
		7914100	7914300									
		7914400	7915400									
		7915700	7915800									
		7916700	7917500									
		7917900	7918200									
		7918300	7918800									
		7918900	7919500									
		7920200	7920300									
		7920500	7920900									
		7921100	7921200									
		7921400	7921500									
		7921900	7923100									
		7923900	7924100									
		7924500	7924700									
		7924800	7926300									
		7926500	7927100									
		7927300	7927900									
		8573200	8573500									
		8583900	8584100									
		9372000	9372100									
		9690400	9690500									
		9765900	9766000									
		11212400	11212500									

		12064900	12065000									
		12065100	12065200									
		12065300	12065400									
		12067200	12067700									
		12067900	12068000									
		12068100	12068300									
		12068400	12068500									
		12068600	12068700									
		12068800	12069300									
		12069500	12069600									
		12069800	12070000									
		12070300	12070400									
		12070700	12070900									
		12071000	12071100									
		12071500	12071600									
		12072000	12072100									
		12072200	12072400									
		12267800	12267900									
		12314000	12314100									
		12314400	12314700									
		12314800	12315000									
		12315100	12315200									
		12315800	12315900									
		12316500	12316900									
		12317200	12317300									
		12317400	12317800									
		12318200	12318600									
		12318800	12318900									
		12319000	12319100									

		12319200	12319300									
		12319600	12319700									
		12320000	12320200									
		12320300	12320400									
		12320800	12320900									
		12321100	12321200									
		12321500	12321600									