HUMAN CHROMOSOME 4 SEQUENCING AND SINGLE NUCLEOTIDE POLYMORPHISM (SNP) ANALYSIS OF AN ACHONDROPLASIA INDIVIDUAL

by

LEE LING SZE

Thesis submitted in fulfillment of the requirements
for the degree of
Master of Science

February 2011

PENJUJUKAN KROMOSOM 4 MANUSIA DAN ANALISIS POLIMORFISME NUKLEOTIDA TUNGGAL (SNP) DARIPADA INDIVIDU ACHONDROPLASIA

oleh

LEE LING SZE

Tesis yang diserahkan untuk
memenuhi keperluan bagi
Ijazah Sarjana Sains

Februari 2011

ACKNOWLEDGEMENTS

First, I thank my supervisor, Prof. Maqsudul Alam, for his continuous support in the Master program. He was always there to listen and to give advice. He was as excited as me when I proposed this project for the first time to him. He taught and guided me different ways to approach a research problem and the need to be persistent to accomplish any goal.

Special thanks goes to my co-supervisors, Prof. Nazalan Najimudin and Dr. Rowani Rawi, who are helping me complete the writing of this dissertation as well as the challenging research that lies behind it. Without their encouragement and constant guidance, I could not have finished this dissertation and project.

Let me also say 'thank you' to the following people at Wellcome Trust Sanger Institute, United Kingdom, Dr. Ng Bee Ling and Willian Cheng, who dedicated their precious time to teach me the techniques on chromosome preparation for flow karyotyping, Dr. Nigel Carter, for giving the opportunity to me to visit and gained fruitful experience in his laboratory, Dr. Chris Detter for helping me on the WGA and Illumina sequencing in this project, and last but not least, Dr. Mike Cariaso, for helping me with the SNP analysis pipeline.

Besides my supervisors, I would also like to thank Dr. Jennifer Saito, who gave useful comments and reviewed my work. I would like to express my gratitude towards all my colleagues and friends in the centre, for the friendship and support, the confidence when I doubted myself, the encouragement and for listening to all my complaints and frustrations.

Last, but not least, I thank my parents and sisters, for unconditional love, support and encouragement to pursue my interests in Science and research.

TABLE OF CONTENTS

CHAPTER 2 – MATERIALS AND METHODS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AVAF	achondroplasia volunteer Asian female
bp	base pair
FGF	fibroblast growth factor
FGFR	fibroblast growth factor receptor
Ig I-III	immunoglobin-like loops I-III
LD	linkage disequilibrium
LINE	long interspersed nucleotide element
MNC	mononuclear cells
PCR	polymerase chain reaction
SADDAN	severe achondroplasia with developmental delay and acanthosis nigricans
SINE	short interspersed nucleotide element
SNP	single nucleotide polymorphism
TDI	thanatophoric dysplasia type I
TDII	thanatophoric dysplasia type II
TM	transmembrane region
TK	tyrosine kinase
WGA	whole genome amplification

Penjujukan Kromosom 4 Manusia dan Analisis Polimorfisme Nukleotida Tunggal (SNP) daripada Individu Achondroplasia

Abstrak

Achondroplasia adalah penyebab paling umum kekerdilan manusia yang beranggota pendek dan mempengaruhi seramai 250,000 orang di seluruh dunia. Penyakit genetik ini menyebabkan pelbagai komplikasi dari segi sosial dan perubatan. Kebanyakan kes achondroplasia berlaku secara rawak dan disebabkan oleh mutasi *de novo*. Gangguan autosomal-dominan ini disebabkan oleh mutasi tunggal dalam gen reseptor jenis 3 faktor pertumbuhan fibroblas (FGFR3). Kajian ini menumpukan pemahaman tentang genetik achondroplasia dengan mengenalpasti SNP daripada kromosom seorang sukarelawan achondroplasia berasal dari Asia. Kaedah pewarnaan kromosom dan penkariotipan aliran bivariat kromosom manusia telah berjaya dioptimumkan. Amplifikasi genom keseluruhan (WGA) telah dilakukan untuk menjana data penjujukan truput tinggi. Analisis data penjujukan dan SNP yang menyeluruh tidak dapat mengenalpasti mutasi yang telah diketahui untuk achondroplasia dan hypochondroplasia. Justeru, kajian ini menunjukkan bahawa penanda gen achondroplasia yang klasik, iaitu gen *fgfr3* bukan satu-satunya penanda dalam kes tertentu ini.

Human Chromosome 4 Sequencing and Single Nucleotide Polymorphism (SNP) Analysis of an Achondroplasia Individual

Abstract

Achondroplasia is the most common cause of short-limbed dwarfism in humans, affecting 250,000 individuals worldwide. This genetic disorder results in various social and medical complications. The majority of achondroplasia cases is sporadic and result from *de novo* mutations. This autosomal-dominant disorder is caused by single nucleotide mutations in the gene encoding the type 3 receptor for fibroblast growth factor (FGFR3). This study focused on understanding the genetic basis of achondroplasia by identifying SNPs from flow-sorted human chromosomes of an achondroplasia volunteer of Asian origin. Chromosome staining and the bivariate flow karyotyping of human chromosomes were successfully optimized. Whole Genome Amplification (WGA) was carried out to generate high-throughput sequencing data. Thorough analysis of the sequence data and SNPs was unable to identify any known mutations of achondroplasia and hypochondroplasia. Thus, it indicates that the classical achondroplasia indicator gene, *fgfr3*, may not be the only indicator in this particular case.

Why am I different?

Living as a shorter person in a world that's designed for the tall people – *Why am I different?* – is the most frequently asked question I always have.

I look different. Everywhere I go, I attract curiosity and I get stared at a lot. As far as I know, I have what I think is an ordinary life. I live with my parents and two sisters. I do not notice the little things that I have to do differently from other people. I felt that I am a normal person, living a normal life. I eat, sleep, breathe, study and get ill, just like everyone else. But why am I still different?

I know that some little people like me have a lot of health problems. Personally, I have walking problems and get more back and joint pain than others my age but this certainly is not enough to stop me to go for sports or activities that I enjoy. Thus, I want to change the lifestyle of a little person, who have more serious health problems than I do, to enable them to lead a normal life like other people.

As Nobel laureate Paul Berg of Stanford University mentioned before "All human disease is genetic in origin." So, how do I investigate the mystery of the genes that made me different and find the answer to my question?

Since the completion of the Human Genome Project, the sequence of the human genome is providing the complete view of the genetic heritage. The human genome, the complete set of human genes, comes in 23 separate pairs of chromosomes. If a human genome is a book, then every human being has a story to tell. Each book comes in 23 chapters, which are called chromosomes. Each chapter contains stories, called genes. Here, I will be telling you the story of one of the chapters in my book, chromosome 4, and focusing one of the stories, the *fgfr3* gene that is related to a one of the best-known genetic diseases, Achondroplasia.

Single-nucleotide polymorphisms (SNPs) are one-base variations in DNA sequence. Each person's genetic material contains a unique SNP pattern that is made up of many different genetic variations. Most SNPs are not responsible for a disease state. Instead, they can often be helpful when trying to find genes responsible for inherited diseases and serve as biological markers for pinpointing a disease on the human genome map. Occasionally, a SNP may actually cause a disease. Therefore, it can be used to search for and isolate the disease-causing gene.

Achondroplasia has been mapped to the tip of the short arm of chromosome 4. So, how can we better understand this genetic disorder? There are two possible ways:

1. Sequence a full human genome and analyze the presence of SNPs, or
2. Study chromosome 4 in-depth and compare the SNP patterns between individuals affected by achondroplasia and individuals unaffected by the genetic disorder.

At the moment, since achondroplasia-associated mutations are already known to be located in chromosome 4, I will first study specifically chromosome 4 to identify SNPs that could be related to the achondroplasia disease family. Now, how can I identify and isolate chromosome 4 from the 23 pairs of chromosomes? One possible way is to use a rapidly developing technique in research and clinical practice, the flow cytometry and sorting instrument. The flow cytometry technique enables us to isolate the desired chromosome from the other chromosomes. Directly after isolation, the flow-sorted chromosomes can be sequenced to determine the nucleotide sequence. As human DNA sequences are 99.9% identical to each other, the 0.1% of variation can provide many clues to many diseases and common illnesses. The identification of such variations can help explore the mystery of achondroplasia.

CHAPTER 1

INTRODUCTION

1.1 Achondroplasia

1.1.1 Overview

Achondroplasia is a Greek word which means "without cartilage formation". This disorder has been present for ages. In fact, people suffering from achondroplasia were used as subjects for art. One of the most famous posterity with the characteristic phenotype of achondroplasia recorded by the artistic community is the portrait of Don Sebastián de Morra (Figure 1.1A), a courtier of Philip V of Spain (Young, 1998), by Velázquez. Achondroplasia was also mentioned in ancient Egypt. Seneb, a Dynasty dwarf, was the chief of the royal wardrobe and priest of the funerary cults of Khufu. A statue still exists of him and it depicts him with his family, including his wife who was of normal stature (Figure 1.1B). Even ancient Egyptian gods such as Bes have been depicted as suffering from achondroplasia (Figure 1.1C) (Kozma, 2006). In fact, throughout history, in the ancient times, many superstitions have been associated with achondroplasia. When a child was born with this condition, it was assumed that it had occurred due to the activities of demons, as a punishment meted out by the gods, or as a result of the movements of the stars and moon.

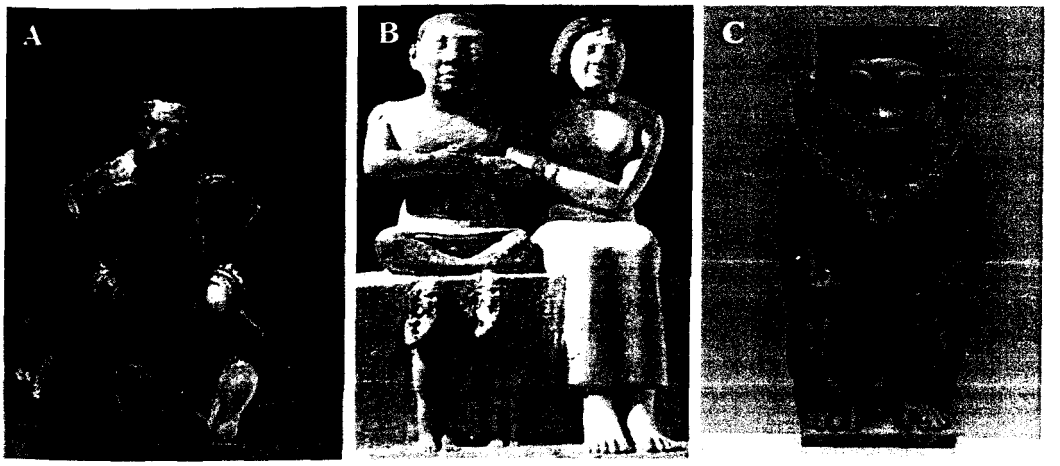

Figure 1.1 History of achondroplasia. **(A)** The portrait of *The Dwarf Don Sebastián de Morra*, by Velázquez. **(B)** Seneb status with his wife and their children. **(C)** Bes statue from Egypt.

Achondroplasia is the most common form of non-lethal skeletal dysplasia. It is the one of the best-known and most frequent cause of short-limbed dwarfism in human beings. Achondroplasia has an incidence rate between one in every 10,000 to one in every 30,000 live births (Oberklaid *et al.*, 1979). More than 85% of achondroplasia cases are sporadic; they are associated with *de novo* mutation (Vajo *et al.*, 2000) and have no familial history. Achondroplasia is estimated to affect more than 250,000 individuals worldwide (Baujat *et al.*, 2008).

The achondroplasia family is characterized by a continuum of severity ranging from mildly affected hypochondroplasia and severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN) to lethal neonatal dwarfism, thanatophoric dysplasia (TD). In individuals with achondroplasia, the skeleton is the primary system involved in the phenotype. All of the disorders in the achondroplasia family of skeletal dysplasias involve some degree of short stature and/or abnormal ossification of bone structures (Vajo *et al.*, 2000). Hypochondroplasia typically present with a mild short stature and a stocky build. TD is much more severe in general and is usually lethal in the prenatal period. SADDAN

refers to a clinical phenotype intermediate in severity between TD and achondroplasia.

Achondroplasia is a disease with shortness in appearance. The characteristics of dwarfism of achondroplasia are so distinctive that they are not difficult to be identified (Castiglia, 1996). Many affected foetuses are recognized in the third trimester of pregnancy. Individuals with achondroplasia are characterized by a long and narrow trunk, short limbs, a large head with a prominent forehead (Figure 1.2A) and a flat depressed nasal bridge (Richette *et al.*, 2008), curved spine (Figure 1.2B), and short hands and fingers with a trident appearance (Figure 1.2C). The average adult height for achondroplasia for both male and female is approximately 4 feet (Baujat *et al.*, 2008; Castiglia, 1996).

Figure 1.2 Features of achondroplasia individual. **(A)** Body disproportion with short limbs, relatively long trunk, and large head (Genetic People, 2009). **(B)** Bending of the spine occurring in the middle and lower back (Nemours Foundation, 2009). **(C)** Trident hands with short fingers (Nemours Foundation, 2009).

Even though the most striking feature of achondroplasia involves cartilage growth, the achondroplasia mutation affects many organ systems (Horton *et al.*, 2007). Many health problems appear at predicted ages, including adulthood. They can be minimized if detected early. Thus, guidelines for individuals with

achondroplasia have been developed in several countries (Horton *et al.*, 2007; Hunter *et al.*, 1998; Trotter and Hall, 2005) to aid physicians in preventive care.

In achondroplasia, various medical complications are consequences of the abnormal linear bone growth. Children with achondroplasia generally have delayed motor milestones such as delays in sitting and walking. It has been reported that tibial bowing, leg and lower back pain are considered to be the hallmarks of achondroplasia (Hunter *et al.*, 1998). Respiratory complications also make a major contribution to achondroplasia, especially in young children (Young, 1998). Sleep dysfunction, including snoring and apnoea, are common in achondroplasia both during daytime and sleep. Apnoea may increase the risk of sudden unexpected death in infants (Hecht *et al.*, 1987). Otitis media or middle ear disease occurs frequently, which will lead to hearing loss if untreated. Speech delay and articulation problems are also recognized complications in achondroplasia. Furthermore, obesity is a major problem in achondroplasia. It can contribute to the potential early cardiovascular mortality in this condition. Occasionally, orthodontic problems such as dental crowding is observed in achondroplasia because of the jaw shortness (Hunter *et al.*, 1998). In addition to all the medical complications, psychological difficulties such as depression are also common among individuals with achondroplasia (Baujat *et al.*, 2008), resulting from the stressful situation required to adapt and cope to the world of taller people.

Even though individuals with achondroplasia run a higher risk for certain health problems, they are able to live a full, normal, and independent life. Most individuals with achondroplasia have normal mental faculties and intelligence (Vajo *et al.*, 2000). Although serious problems may arise during infancy, they affect only 5% to 10% of infants with achondroplasia (Trotter and Hall, 2005).

In addition, individuals with achondroplasia can also lead a productive life. Sexual development and fertility seems to be normal in achondroplasia-affected women who opt for childbearing (Horton *et al.*, 2007; Richette *et al.*, 2008). The diagnosis of achondroplasia in the foetus is made with certainty when one or both parents have this condition. Fifty percent of the offspring of individuals with achondroplasia will be affected (Baujat *et al.*, 2008). When both parents have typical achondroplasia, their children with homozygous achondroplasia generally do not survive beyond a few weeks or possibly a few months (Castiglia, 1996).

1.1.2 Human chromosome 4

Chromosome 4 is one of the 23 pairs of chromosomes in humans. Chromosome 4 contains approximately 190 million base pairs and comprises 6.5 percent of the total human genomic DNA (Gusella *et al.*, 1986). Hillier *et al.* (2005) have identified 796 protein-coding genes and 778 pseudogenes on chromosome 4. Chromosome 4 contains the highest percentage of the long interspersed nucleotide element (LINE) content across all autosomes. However, the short interspersed nucleotide element (SINE) content is lower than the autosomal average. One of the highest (G+C) content windows in the genome is also found on chromosome 4. Hillier *et al.* (2005) also identified 1,004 CpG islands in chromosome 4 (5.4 per Mb), analyzed based on 186 million base pairs, each with an average length of approximately 800 bp. Chromosome 4 has the lowest density of predicted CpG islands and the lowest average recombination rate of any of the chromosomes (Hillier *et al.*, 2005).

Identifying genes on each chromosome is an active area of genetic research. As researchers use different approaches to predict the number of genes on each

chromosome, the estimated number of genes varies. Some of the famous diseases related to genes located on chromosome 4 are Huntington disease, Ellis-van Creveld syndrome, and Parkinson disease.

The gene responsible for achondroplasia was genetically mapped to the short arm of chromosome 4, 4p16.3 (Le Merrer *et al.*, 1994; Velinov *et al.*, 1994). Significantly, it was mapped very close to another elusive disease gene locus, the Huntington disease. Together with the discovery of the gene causing Huntington's disease, increased interest was generated towards this chromosome (Figure 1.3).

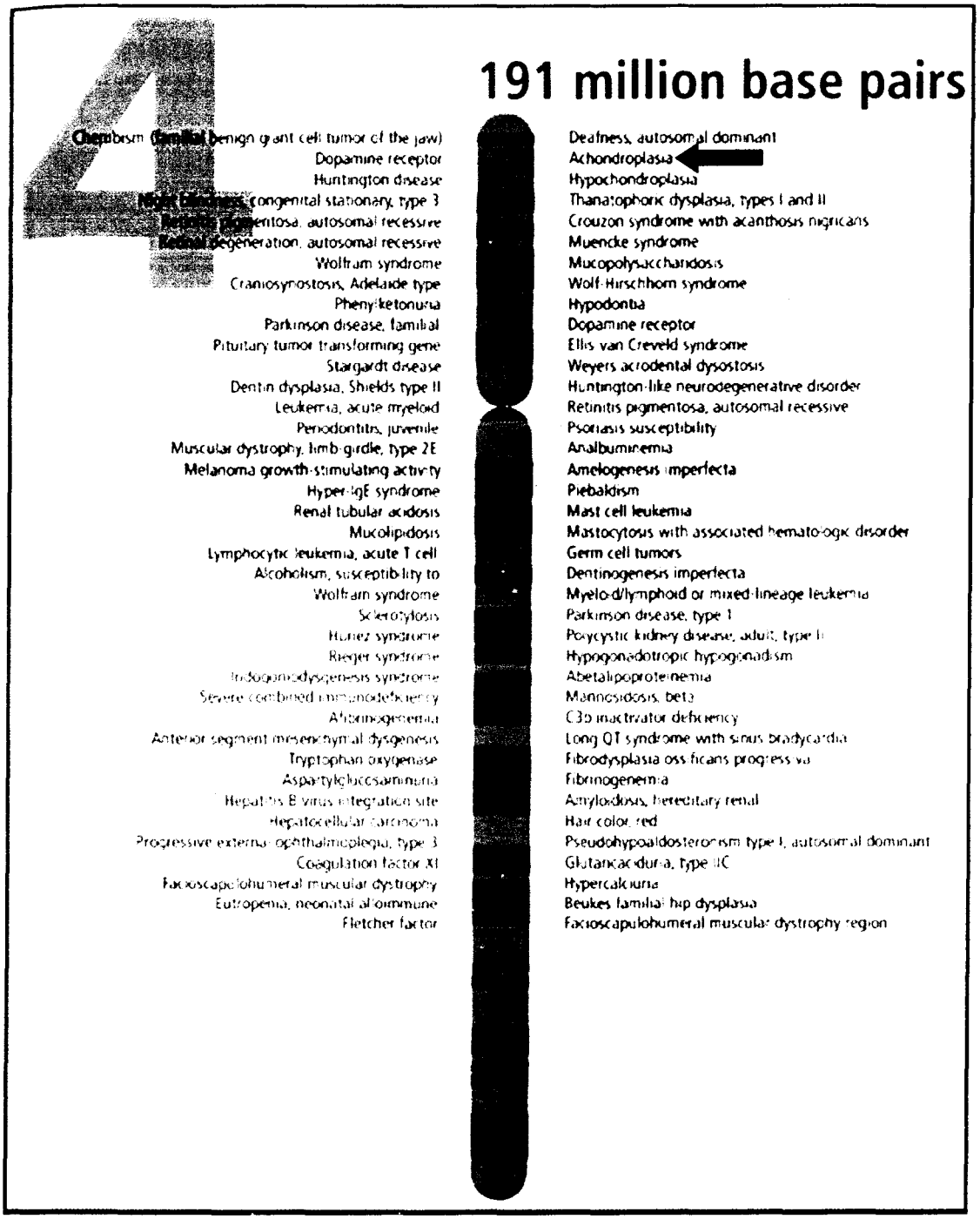

Figure 1.3 Human chromosome 4 and diseases mapped to the chromosome. Adapted from U.S. Department of Energy Genome Program (DNARSS.com, 2009).

1.1.3 Fibroblast growth factor receptor 3 (FGFR3)

The gene causing achondroplasia was discovered by Dr. John Wasmuth (Shiang *et al.*, 1994). While working with his colleagues, Wasmuth discovered that a mutation in the fibroblast growth factor receptor 3 (*fgfr3*) gene caused this autosomal-dominant disorder. In 1993, Keegan *et al.* reported that the *fgfr3* gene localizes to human chromosome 4p16.3, confirming the existence of *fgfr3* genes (Keegan *et al.*, 1993). The identified causative mutations in *fgfr3* responsible for achondroplasia showed that a mutation in a transmembrane domain of this fibroblast growth factor receptor results in a skeletal growth defect (Rousseau *et al.*, 1994).

FGFR3 plays an important role in long bone development. FGFR3 belongs to the fibroblast growth factor receptor family. FGFR3 is one of the four FGFR members (FGFR 1-4), which share a common organization comprising three extracellular immunoglobin-like loops (Ig I-III), a single hydrophobic transmembrane region (TM), and two cytoplasmic tyrosine kinase (TK) subdomains TK1 and TK2 (Figure 1.4) (Schlessinger, 2000). The *fgfr3* gene contains an open reading frame of 2905 nucleotides and consists of 19 exons and 18 introns (Table 1.1).

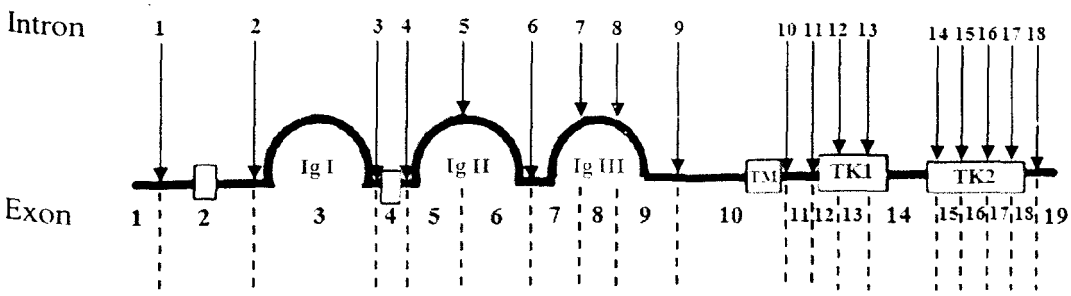

Figure 1.4 Structure and organization of the human *fgfr3* gene. Numbers above the arrows and the vertical dashed lines indicate the positions of intron and exon sequences, respectively. Sizes of both introns and exons are not drawn to scale (Wuchner *et al.*, 1997).

Table 1.1 Exon and intron sizes of the human *fgfr3* gene (Wuchner *et al.*, 1997)

Exon	Exon size (bp)	Intron	Intron size (bp)
1	171	1	368
2	211	2	5210
3	270	3	223
4	66	4	1554
5	170	5	83
6	124	6	91
7	191	7	888
8	151	8	627
9	145	9	492
10	191	10	303
11	146	11	385
12	122	12	82
13	111	13	80
14	191	14	110
15	123	15	83
16	71	16	218
17	138	17	145
18	106	18	181
19	207	-	-

FGFR3 is one of many important local physiological regulators of linear bone growth (Horton and Lunstrum, 2002). Studies suggested that FGFR3 was a negative regulating factor of endochondral ossification (Deng *et al.*, 1996). It binds with the fibroblast growth factors (FGFs). From the 22 known FGF ligands, the exact physiological ligands for FGFR3 is not known, although FGFs 2, 4, 9, and 18 are probably the best candidates based on the distribution of expression and ability to bind and activate FGFR3 (Horton *et al.*, 2007). The developmental expression pattern of FGFR3 suggests that this protein plays a significant role in skeletal/bone development (Vajo *et al.*, 2000). Direct evidence for the discovery that mutations in the coding sequences of *fgfr3* gene cause bone abnormalities in humans was reported by Rousseau *et al.* (1994).

1.1.4 Genetics of achondroplasia

The clinical spectrum of the achondroplasia family of disorders is caused by different mutations in *fgfr3*. In most cases of achondroplasia, the genetic abnormality is due to a mutation located within a critical region of *fgfr3* (Richette *et al.*, 2008). It has been demonstrated that all new mutations occur on the mutated allele from a paternal origin, suggesting increased mutability of *fgfr3* by an increased paternal age at the time of conception (Wilkin *et al.*, 1998). Almost all individuals with achondroplasia are caused by one of two point mutations in the gene for *fgfr3* (Young, 1998). The mutations are a G-to-A transition (G1138A) and a G-to-C transition (G1138C) in nucleotide 1138 of the *fgfr3* gene (Bellus *et al.*, 1995). Both mutations result in the same amino acid substitution (Gly380Arg) in the transmembrane domain of FGFR3 (Figure 1.5) (Baujat *et al.*, 2008). The relatively high incidence of achondroplasia suggests that nucleotide 1138 of the *fgfr3* gene is the most mutable nucleotide described so far in the human genome (Vajo *et al.*, 2000).

Hypochondroplasia is caused by mutations in tyrosine kinase domain 1 (Asn540Lys, Asn540Thr, or Asn540Ser) and tyrosine kinase domain 2 (Lys650Asn and Lys650Gln). Additional substitutions occur at positions 538 (Ile538Val), 278 (Tyr278Cys), and 84 (Ser84Leu) (Grigelioniene *et al.*, 2000). Several mutations in the extracellular domain or the stop codon (Vajo *et al.*, 2000) are associated with thanotophoric dysplasia type I, while a mutation in tyrosine kinase domain 2 (Lys650Glu) is associated with thanatophoric dysplasia type II, which is also lethal but less severe (Figure 1.5 and Table 1.2).

The findings in individuals with achondroplasia prompted the search for *fgfr3* mutations in other disorders considered related to achondroplasia (Vajo *et al.*, 2000).

For instance, a G-to-A transition in nucleotide 1172 has been identified in individuals with Crouzon syndrome with Acanthosis Nigricans, resulting in an Ala391Glu (A391E) substitution in the transmembrane domain. On the other hand, Muenke syndrome has a Pro250Arg (P250R) amino acid substitution, caused by a C-to-G transition at position 749 of the coding cDNA sequence.

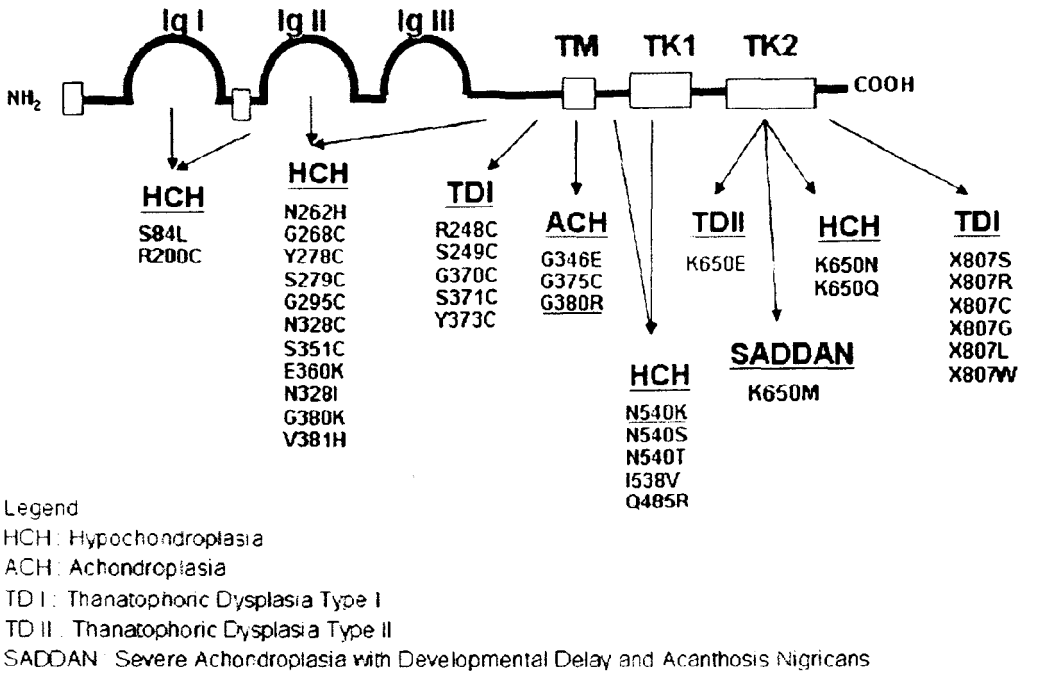

Figure 1.5 FGFR3 mutations identified in chondrodysplasias (Baujat *et al.*, 2008). Adapted from Figure 5, page 12, Baujat *et al.*, 2008.

Table 1.2 Nucleotide transitions and amino acid substitutions in achondroplasia family

Achondroplasia family	Mutation	Substitution resulted	References
Achondroplasia	G1138A/G1138C	Gly380Arg	(Rousseau *et al.*, 1994)
	T1130G	Leu377Arg	(Heuertz *et al.*, 2006)
	G1123T	Gly375Cys	(Chen *et al.*, 1999)
	G1037A	Gly346Glu	(Baujat *et al.*, 2008)
Hypochondroplasia	C1659A/C1659G	Asn540Lys	(Bellus *et al.*, 2000)
	A1658C	Asn540Thr	(Grigelioniene *et al.*, 2000)
	A1658G	Asn540Ser	(Baujat *et al.*, 2008)
	A1651G	Ile538Val	(Grigelioniene *et al.*, 2000)
	G1950T/G1950C	Lys650Asn	(Bellus *et al.*, 2000)
	A1948C	Lys650Gln	(Bellus *et al.*, 2000)
	A831T	Ser279Cys	(Heuertz *et al.*, 2006)
	A829G	Tyr278Cys	(Heuertz *et al.*, 2006)
	C251T	Ser84Leu	(Heuertz *et al.*, 2006)
	G801T	Gly268Cys	(Heuertz *et al.*, 2006)
	A783C	Asn262His	(Heuertz *et al.*, 2006)
	C597T	Arg200Cys	(Heuertz *et al.*, 2006)
	A983T	Asn328Ile	(Bellus *et al.*, 2000)
	G879T	Gly295Cys	(Baujat *et al.*, 2008)
	C1052G	Ser351Cys	(Baujat *et al.*, 2008)
	G1081A	Glu360Lys	(Baujat *et al.*, 2008)
Thanatophoric dyplasia (TD) Type I	C742T	Arg248Cys	(Rousseau *et al.*, 1996)
	C746G	Ser249Cys	(Rousseau *et al.*, 1996)
	A1111T	Ser371Cys	(Rousseau *et al.*, 1996)
	T2458G	Stop807Gly	(Rousseau *et al.*, 1996)
	T2458A	Stop807Arg	(Rousseau *et al.*, 1996)
	A2460T	Stop807Cys	(Rousseau *et al.*, 1996)
	G1108T	Gly370Cys	(Rousseau *et al.*, 1996)
	A1118G	Tyr373Cys	(Rousseau *et al.*, 1996)
Thanatophoric dyplasia (TD) Type II	A1948G	Lys650Glu	(Tavormina *et al.*, 1999)
Severe achondroplasia with developmental delay and Acanthosis nigricans (SADDAN)	A1949T	Lys650Met	(Tavormina *et al.*, 1999)

1.1.5 Single nucleotide polymorphisms (SNPs)

SNP (pronounced 'S' 'N' 'P' or 'SNiP') stands for Single Nucleotide Polymorphism. SNPs are the most common and abundant form of genetic variation in humans (Taillon-Miller *et al.*, 1998). Simply put, they are single base pair positions in genomic DNA at which different sequence alternatives exist in normal individuals in some populations. SNPs commonly occur at a rate greater than 1% in a given population. About 90% of all the sequence variation recorded in the human genome is due to SNPs (Collins *et al.*, 1997). In the human genome, over 3 million SNPs have been identified (Brookes, 1999). The typical frequency in which a single base differs in the genomic DNA from two equivalent chromosomes is one per kilobase pair of sequence (Taillon-Miller *et al.*, 1998). It is estimated that over 99% of the human genome sequence is conserved across various populations.

One of the most frequently reported mutations found in the majority of achondroplasia-affected individuals is the G-to-A transition (G1138A) in nucleotide 1138 of the *fgfr3* gene (Bellus *et al.*, 1995), resulting in the amino acid substitution in the transmembrane domain of FGFR3 (Gly380Arg).

Genetic factors such as SNPs may not directly cause disease but confer susceptibility or resistance to a disease or determine the severity or progression of disease (Collins *et al.*, 1998). SNPs can help determine the likelihood that someone will develop a particular disease. They can have a major impact on the way humans respond to disease and environmental insults such as bacteria, viruses, toxins, chemicals, drugs, and other therapies. This makes SNPs of great value for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs are also evolutionarily stable; they do not change significantly from generation to generation, making them easier to follow in population studies.

SNPs occur in both coding regions as well as non-coding regions. Most SNPs fall in the non-coding region of the human genome, presumably due to lower selection pressure. The frequency of SNPs in the coding region is observed to be 4-fold lower than in non-coding regions (Collins *et al.*, 1997) because such sequence alterations can result in changing the transcript and hence the corresponding protein. Therefore, the SNPs in these regions have a direct capability to significantly impact the shape, structure, or a critical residue in the protein which might ultimately result in aberrant function of the protein and result in a disease.

Of the SNPs that are near or in a gene, their effect on function is difficult to determine. SNPs are generally classed by genomic location. SNPs can fall within the coding regions, regulatory regions, in exons, or within introns. Non-synonymous SNPs (nsSNP) alter the amino acid sequence of the protein product through amino acid substitution. A variant may also affect the expression or translation of a gene product, either by interrupting a regulatory region or by interfering with normal splicing and mRNA function. This can include regulatory SNPs (rSNP), synonymous SNPs (sSNP), and intronic SNPs (iSNP). The two types of variation that are usually studied are polymorphisms with known phenotype and phenotypically annotated or disease-associated variation. Human mutations are often inferred to be disease-associated (Mooney, 2005).

Most SNPs do not directly result in disease since most diseases are due to aberrant errors in a number of genes, such as in cancer, heart diseases, and diabetes (Houlston and Peto, 2004; Pharoah *et al.*, 2004). However, there are some diseases that have been linked to a single gene, such as Huntington's disease, haemophilia, or sickle cell anemia. Variation does not occur randomly across genetic sequences and

often occurs in hotspots (Benzer, 1961). It is likely that selection has played a role in the evolution of human genetic variation (Akey *et al.*, 2002; Fay *et al.*, 2001).

1.1.6 Treatment

A single nucleotide change in the human genome can make such a big appearance difference in a human being. The mortality rate in individuals with achondroplasia is higher than the general population, particularly during childhood. The cause of this increased mortality rate in young children is attributable to severe cervicomedullary compression (Hunter *et al.*, 1998). Until today, current therapies for the short stature in achondroplasia are still debated as there is no treatment that exists to reverse the genetic abnormality of achondroplasia. Administrations of growth hormone and surgical limb-lengthening procedures have been proposed (Seino *et al.*, 2000). Human growth hormone therapy has been used as a treatment for the short stature in children with achondroplasia. Although there has been some increase in growth rate reported, long-term benefits are not conclusive. Thus, most experts do not recommend such treatment for achondroplasia (Horton *et al.*, 2007). Surgical limb-lengthening is another approach using several surgical and orthopaedic appliances. It involves breaking bones, followed by slow stretching during the healing process (Horton *et al.*, 2007). However, this procedure remains arduous with a high risk of infection, joint and soft tissue damage, and may result in a poorer quality of life (Aldegheri *et al.*, 1988).

1.2 Flow cytometry

1.2.1 Overview of flow cytometry

Flow cytometry is an extremely powerful and exciting technology involving the analysis of fluorescence and light scatter properties of microscopic particles at high speed. It allows the individual measurement of physical and chemical characteristics of particles as they pass one by one through a light source. The method was originally developed for the analysis of blood cells. Currently, most flow cytometers are used to evaluate human cells stained with various dyes and labelled with a variety of antibodies. The range of applications has continued to increase and encompass the analysis of ploidy, cell cycle kinetics, and the presence of specific antigens (Dolezel *et al.*, 2004). Flow cytometers and sorters have become a widespread and vital resource in the biological sciences and beyond. It is a process that allows the physical separation of a cell or particle of interest from a heterogeneous population.

1.2.2 Flow cytogenetics

Early discussions about sequencing the entire human genome were considered credible in large part due to the ability to flow sort, with high purity, each of the human chromosomes. High-purity sorting made it possible to clone and produce chromosome-specific libraries suitable for sequencing (Cram *et al.*, 2004).

At first, it seems improbable that the founders of flow cytometry thought of analyzing chromosomes with these instruments. Yet, the meeting of flow cytometry and cytogenetics gave rise to a whole new area of research called flow cytogenetics. Flow cytogenetics describes the application of flow cytometry for analysis and

sorting of mitotic chromosome classification and purification (Cram *et al.*, 2002). Flow cytogenetics has contributed significantly to the progress in many areas of genome analysis and mapping as well as underpinning the sequencing of the human genome (Dolezel *et al.*, 2004).

The underlying principles of flow cytogenetics are relatively straightforward. The chromosomes in an aqueous suspension are constrained to flow in a single file within a fluid stream and past a narrow beam of excitation light. During the short time each chromosome is in the light beam, the light is scattered and the molecules of fluorochrome bound to the chromosomes are excited.

In flow cytogenetics, a large number of fluorescent dyes are capable of interacting with DNA. When such dyes are used individually to stain cells or chromosomes, their fluorescence can be influenced not only by the amount of DNA present but by the DNA base composition (Latt *et al.*, 1979). The persistent problem was the inability to resolve all chromosomes within a karyotype due to the similarity of relative DNA content (Dolezel *et al.*, 2004). This was overcome by improving the existing procedures for chromosome isolation and by staining the chromosome preparation with two dyes differing in base pair preference, such as Hoechst 33258 and Chromomycin A3 (Latt *et al.*, 1979). Although various other approaches were introduced to improve chromosome discrimination, bivariate analysis using Hoechst and Chromomycin has become the gold standard for chromosome analysis using flow cytometry/flow karyotyping in human and animals (Dolezel *et al.*, 2004).

1.2.3 Flow karyotype

Flow karyotyping provides precise information about chromosome properties, such as DNA content for several hundred thousand chromosomes (Cram *et al.*, 2002). A flow karyotype is the distribution of relative fluorescence intensity of individual chromosomes or groups of chromosomes of similar relative DNA content. This opened an exciting avenue towards the purification of individual chromosomes by flow sorting (Dolezel *et al.*, 2004). Flow karyotyping requires isolation of intact metaphase chromosomes, staining the chromosome suspension with a fluorescent tag, and rapid quantitative analysis in a flow cytometer.

Applications of univariate (one colour) flow karyotype analysis include determining and monitoring karyotype instability, variation in the frequency of a chromosome type, chromosomal polymorphisms, and chromosome rearrangements. For univariate flow karyotyping, chromosome discrimination is based on the amount of fluorescent dye bound to the chromosome. Many of the fluorochromes used for flow karyotyping bind only to nucleic acids so that discrimination is largely based on total DNA content.

Bivariate flow karyotyping, where chromosome classification is based on two fluorochromes, was developed to take advantage of the fact that some dyes like Hoechst 33258 and Chromomycin CA3 bind preferentially to adenine-thymine (AT) or guanine-cytosine (GC) rich DNA, respectively. This pair of fluorochromes allows classification of chromosomes according to DNA content and DNA base composition. Figure 1.6 shows a typical bivariate human flow karyogram and Table 1.3 shows the size of all the human chromosomes and the estimation of known protein-coding genes in each chromosome.

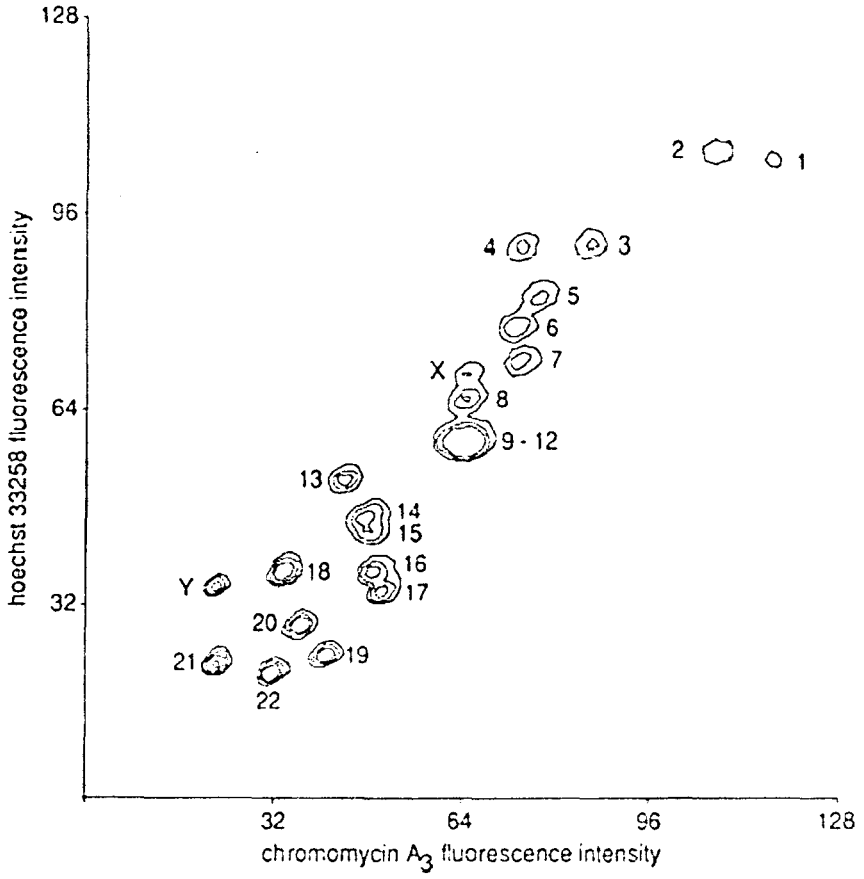

Figure 1.6 A typical bivariate flow karyogram of a normal human male cell (Cram *et al.*, 2002). Adapted from Figure 5, page 30, Cram *et al.*, 2002.

Table 1.3 Human chromosome sizes and an estimate of the number of known protein-coding genes of each chromosome

Chromosome	Size (bp)	Number of known protein-coding genes
1	249,250,621	2029
2	243,199,373	1230
3	198,022,430	1055
4	**191,154,276**	**796**
5	180,915,260	867
6	171,115,067	1022
7	159,138,663	973
8	146,364,022	755
9	141,213,431	806
10	135,534,747	767
11	135,006,516	1352
12	133,851,895	1051
13	115,169,878	324
14	107,349,540	633
15	102,531,392	671
16	90,354,753	907
17	81,195,210	1184
18	78,077,248	287
19	59,128,983	1456
20	63,025,520	551
21	48,129,895	235
22	51,304,566	445
X	155,270,560	833
Y	59,373,566	48

Notes: Chromosome sizes and number of known protein-coding genes according to GRCh37 from Ensembl (Ensembl, 2010).

1.2.4 Chromosome sorting

Chromosome isolation consists of freeing individual chromosomes from mitotic cells and stabilizing their structure. Staining reactions are designed to label a mixture of chromosome types so that one chromosome type is distinguished from another. The ultimate goal is to resolve each chromosome type from any given species. Chromosome purification by sorting requires the highest possible discrimination of chromosome types from one another and from chromosomal debris and clumps. In the case of chromosomes isolated from human cells, this means

resolving 23 populations when using cells of female origin (22 autosomes and X chromosome) and 24 populations in cells of male origin (22 autosomes, X and Y chromosomes). The ability to resolve all chromosome types from any mammalian species usually depends upon differences in inter-chromosomal DNA content, either total DNA content or base pair ratios, and instrumental resolution. Chromosome sorting is used to identify chromosome types in a flow karyotype and has been extensively used for gene mapping, cloning, and molecular characterization of normal and rearranged chromosomes (Cram *et al.*, 2002).

Chromosome sorting and analysis played a major role in the early stages of the human genome program. New genome-related applications continue to evolve in the areas of genomics and proteomics. Five major areas of application have developed: flow cytogenetics, construction of chromosome specific libraries, bead-based assays for detection of single nucleotide polymorphisms (SNPs), DNA fragment analysis, and single molecule DNA sequencing. Clinical applications in flow cytogenetics have evolved around the ability to detect and sort aberrant chromosomes due to translocation, deletion or addition. In particular, the identification of translocations by the application of chromosome-specific probes derived from sorted chromosomes. The single largest application of chromosome sorting has been the creation of chromosome-specific libraries. Human chromosome-specific libraries provided the initial starting material that was used in the early stages of the human genome project. The availability of human libraries constructed from a single human chromosome simplified the project by being able to assign and map DNA sequences known to have come from a single chromosome type. New developments in bead-based assays, DNA fragment analysis, and single molecule DNA sequencing further demonstrate the versatility of flow cytometry to measure and analyze genetic changes at the

molecular level. Bead-based flow cytometric assays are being used to detect single nucleotide polymorphisms (SNPs). DNA fragments have been analyzed in specialized flow cytometers capable of photon counting. All the necessary components of single molecule DNA sequencing have been demonstrated using specialized flow cytometers to rapidly sequence very long DNA segments.

1.3 Bioinformatics

Currently, there are many human SNPs databases available to compare and analyze SNPs.

1.3.1 Database on human SNPs / SNP analysis

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has led to an absolute requirement for computerized methods to store, organize, and index the data and specialized tools to view and analyze the data. With the completion of the human genome project in 2002 (Lander *et al.*, 2001) and further refinement over the past few years (International Human Genome Sequencing Consortium, 2004), a complete catalogue of all the human genes, their sequences, and locations within the genome is currently available.

Over the past decade, considerable effort has been placed on understanding the genetic changes that give rise to the molecular effects that cause diseases and phenotypes (Mooney, 2005). These efforts have given rise to many databases, web resources, and tools for prioritizing candidate SNPs or hypothesizing the molecular causes of genetic disease, with most of the focus on human annotations (Mooney *et*

al., 2010). Functional bioinformatics approaches have been applied to the analysis of disease-associated mutations. One of the difficulties in analysing disease-associated mutations is that it is very difficult to obtain a set of neutral alleles for comparison (Mooney, 2005).

There are now many databases that provide access to SNP or disease mutation data. Most SNP data is eventually deposited in the primary SNP database, The Single Nucleotide Polymorphism database (dbSNP), which contains more than 5,000,000 validated human SNPs. There are also many disease-associated and genotype-phenotype polymorphisms databases available such as the Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2000), Swiss-Prot (Boeckmann *et al.*, 2003), the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2003), HGVBase (Fredman *et al.*, 2004), the Pharmacogenetics Knowledge Base (PharmGKB) (Altman, 2007), and database of Genotype and Phenotype (dbGAP) (Mailman *et al.*, 2007).

Many resources now annotate variation data with functional information. Information about whether variants occur near a gene, in a coding region, in an exon, in an intron, or upstream or downstream of the gene are relatively direct using several genome resources. The NCBI databases, such as dbSNP and OMIM (Wheeler *et al.*, 2001), and Ensembl (Hammond and Birney, 2004) provide visualisation access and some annotations related to function, based on experimental data.

In order to predict genes that are likely to cause or be associated with disease, a recent disease gene prioritization tool is FitSNPs (Chen *et al.*, 2008). The tool is claimed to provide a new way to distinguish disease-associated genes from false positives in genome-wide association studies. GeneSeeker (van Driel *et al.*, 2005)

produces a list of candidate disease genes based on cytogenetic localization and expression/phenotypic data from various human and mouse databases. Transcriptomics of OMIM (Rossi *et al.*, 2006) identifies candidate genes involved in inherited diseases. Gentrepid (George *et al.*, 2006) aims to improve some of the existing methods for candidate gene prediction by using structural bioinformatics and system biology approaches such as domain comparison, pathways, and protein-protein interaction data.

The useful approach to undertake for identification of functional sites near genetic variation data is to identify functional features that reside on or near the site of variability. Several SNP or mutation specific databases have been developed that provide a variety of genomic annotations. There are now many resources for prediction of functional SNPs. Many bioinformatic tools are available to predict functional sites in protein sequences and structures and several resources annotate SNPs with transcript level features (Mooney *et al.*, 2010). One challenge in the identification of human functional SNPs is that many SNPs may be in linkage disequilibrium (LD) with each other. That is, pairs or groups of SNPs may be highly correlated within a population, preventing accurate statistical identification of the causal element (Hudson, 2003). There are several SNP browsing tools that can identify features in the promoter region and relate that information to SNPs that are present upon them. These include the NCBI genome database (Pruitt and Maglott, 2001), SNP@Promoter (Kim *et al.*, 2008), the SNP Function Portal (Wang *et al.*, 2006), and PupaSuite (Conde *et al.*, 2004).

An excellent resource for visualisation of SNP locations and other genome annotations is GoldenPath, the UCSC Genome Browser and genome assembly (Kent *et al.*, 2002). The database is completely in the public domain. Another powerful