View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Publications from Karolinska Institutet

Institutionen för Biovetenskaper och Näringslära

Application of next generation sequencing in genetic and genomic studies

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i 9Q Månen, Alfred Nobels Allé 8, Karolinska Institutet, Huddinge

Fredagen den 9 september, 2016, kl 09.30

av
Jingwen Wang
MSc.

Huvudhandledare:
Dr. Hong Jiao
Karolinska Institutet
Institutionen för Biovetenskaper och Näringslära

Bihandledare:
Professor Juha Kere
Karolinska Institutet
Institutionen för Biovetenskaper och Näringslära

Dr. Thomas Svensson
Chalmers Tekniska Högskolan
Institutionen för Biologi och Bioteknik

Fakultetsopponent:
Professor Sampsa Hautaniemi
Helsinfors Universitet
Medicinska Fakulteten

Betygsnämnd:
Professor Mauno Vihinen
Lunds Universitet
Institutionen för Experimentell Medicinsk Vetenskap

Docent Olof Emanuelsson
Kungliga Tekniska Högskolan
Skolan för Bioteknologi

Docent Ingrid Kockum
Karolinska Institutet
Institutionen för Klinisk Neurovetenskap

From Department of Biosciences and Nutrition
Karolinska Institutet, Stockholm, Sweden

APPLICATION OF NEXT GENERATION SEQUENCING IN GENETIC AND GENOMIC STUDIES

Jingwen Wang

王靜文

Stockholm 2016

All previously published papers were reproduced with permission from the publisher.
Front cover shows the sequence of zebrafish new transcripts discovered and validated in this study. Cover page illustration by Zidong Lin.
Published by Karolinska Institutet.
Printed by E-print AB
© Jingwen Wang, 2016
ISBN 978-91-7676-321-6

Application of next generation sequencing in genetic and genomic studies

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Jingwen Wang

Principal Supervisor:
Dr. Hong Jiao
Karolinska Institutet
Department of Biosciences and Nutrition

Co-supervisor(s):
Professor Juha Kere
Karolinska Institutet
Department of Biosciences and Nutrition

Dr. Thomas Svensson
Chalmers Unversity of Techonology
Department of Biology and Biological Engineering

Opponent:
Professor Sampsa Hautaniemi
University of Helsinki
Faculty of Medicine

Examination Board:
Professor Mauno Vihinen
Lund University
Department of Experimental Medical Science

Docent Olof Emanuelsson
KTH Royal Institute of Technology
School of Biotechnology

Docent Ingrid Kockum
Karolinska Institutet
Department of Clinical Neuroscience

As you do not know the path of the wind,
or how the body is formed in a mother's womb,
so you cannot understand the work of God, the Maker of all things.

-- Ecclesiastes 11:5 (NIV)

獻給我摯愛的父母

感謝他們永不止息的愛

ABSTRACT

Genetic variants that spread along the human genome play vital roles in determining our traits, affecting development and potentially causing disorders. Most common disorders have complex underlying mechanisms involving genetic or environmental factors and the interaction between them. Over the past decade, genome-wide association studies (GWAS) have identified thousands of common variants that contribute to complex disorders and partially explain the heritability. However, there is still a large portion that is unexplained and the missing heritability may be caused by several factors, such as rare or low-frequency variants with high effect that are not covered by GWAS and linkage analysis. With the development of next generation sequencing (NGS), it is possible to rapidly detect large amount of novel rare and low-frequency variants simultaneously at a low cost. This new technology provides vast information on studying the association of genetic variations and complex disorders. Once the susceptibility gene is mapped, model organisms such as zebrafish (*Danio rerio*) are popular for further investigating the possible function of disease-associated gene in determining the phenotype. However, the genome annotation of zebrafish is not complete, which affects the characterization of gene functions. Accordingly, high-throughput RNA sequencing can be employed for identifying new transcripts.

In our studies, pooled DNA samples were used for whole genome sequencing (WGS) and exome sequencing. In **Paper I**, we evaluated minor allele frequency (MAF) estimates using three variant detection tools with two sets of pooled exome sequencing and one set of pooled WGS data. The MAFs from the pooled sequencing data demonstrated high concordance (r = 0.88-0.94) with those from the individual genotyping data. In **Paper II**, exome sequencing implementing pooling strategy was performed on 100 idiopathic scoliosis (IS) patients for mapping susceptibility genes. After validating 20 candidate single nucleotide variants (SNVs), we did not find associations between them and IS. However, the previously reported common variant rs11190870 near *LBX1* was validated in a large Scandinavian cohort. **In Paper III**, we analyzed WGS of pooled DNA samples performed on 19 affected individuals who shared a phenotype-linked haplotype in a dyslexic Finish family. Two of the individuals were sequenced for the whole genome individually as well. The screen for causative variants was narrowed down to a rare SNV, which might affect the binding affinity of *LHX2* that regulated dyslexia associated gene *ROBO1*. In **Paper IV**, RNA sequencing (RNA-seq) data were analyzed for identifying novel transcripts in zebrafish early development using an in-house pipeline. We discovered 152 novel transcribed regions (NTRs), validated more than 10 NTRs and quantified their expression in early developmental stages.

In our studies, we evaluated and applied a pooling approach for identifying variants susceptible to disease using high-throughput DNA sequencing. Based on RNA sequencing data, we provided new information for genome annotation on model organism zebrafish, which is valuable for studying the function of disease causative genes. In summary, the whole series of studies demonstrate how NGS can be applied in studying the genetic basis of complex disorders and assisting in follow-up functional studies in model organisms.

LIST OF SCIENTIFIC PAPERS

I. **Wang J**, Skoog T, Einarsdottir E, Kaartokallio T, Laivuori H, Grauers A, Gerdhem P, Hytönen M, Lohi H, Kere J, Jiao H.
Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples.
Manuscript submitted to *Sci Rep*. and under review

II. Grauers A[†], **Wang J**[†], Einarsdottir E, Simony A, Danielsson A, Åkesson K, Ohlin A, Halldin K, Grabowski P, Tenne M, Laivuori H, Dahlman I, Andersen M, Christensen SB, Karlsson MK, Jiao H, Kere J, Gerdhem P.
Candidate gene analysis and exome sequencing confirm *LBX1* as a susceptibility gene for idiopathic scoliosis.
Spine J. 2015 Oct 1;15(10):2239-46. doi: 10.1016/j.spinee.2015.05.013.

III. Massinen S, **Wand J**, Laivuori K, Bieder A, Paez IT, Jiao H, Kere J.
Genomic sequencing of a dyslexia susceptibility haplotype encompassing *ROBO1*.
J Neurodev Disord. 2016 Jan 27;8:4. doi: 10.1186/s11689-016-9136-y.

IV. **Wang J**, Vesterlund L, Kere J, Jiao H.
Identification of novel transcribed regions in zebrafish (*Danio rerio*) using RNA-sequencing
Manuscript submitted to *PLoS One* and under review

[†] Equal contribution to the work

Other publications not involved in this thesis

I. Smialowska A, Djupedal I, **Wang J**, Kylsten P, Swoboda P, Ekwall K.
RNAi mediates post-transcriptional repression of gene expression in fission yeast Schizosaccharomyces pombe.
Biochem Biophys Res Commun. 2014 Feb 7;444(2):254-9

II. Kaartokallio T[†], **Wang J**[†], Heinonen S, Kajantie E, Kivinen K, Pouta A, Gerdhem P, Jiao H, Kere J, Laivuori H.
Exome sequencing in pooled DNA samples to identify maternal pre-eclampsia risk variants.
Sci Rep. 2016. In press

[†] Equal contribution to the work

CONTENTS

1 Background 1
  1.1 Genetic variations 1
    1.1.1 Different types of genetic variations 1
    1.1.2 Effects 2
  1.2 Disorders 3
    1.2.1 Monogenic disorders 3
    1.2.2 Complex disorders 3
  1.3 Gene mapping in disorders 4
    1.3.1 Linkage analysis 5
    1.3.2 Association studies 5
  1.4 Missing heritability of complex disorders 6
  1.5 Functional studies following identification of disease-causative genes 7
    1.5.1 *In silico* 8
    1.5.2 *In vivo* – model organisms 8
2 Introduction 9
  2.1 NGS platforms 9
    2.1.1 Illumina 10
    2.1.2 SOLiD 11
    2.1.3 Complete Genomics 11
  2.2 NGS application 11
    2.2.1 Whole genome sequencing and exome sequencing 11
    2.2.2 RNA sequencing 14
  2.3 Bioinformatic analysis of NGS data 15
3 Aims 16
4 Materials and methods 17
  4.1 Study subjects and materials 17
    4.1.1 Idiopathic scoliosis (IS) case-control cohorts 17
    4.1.2 Preeclampsia (PE) case-control cohorts and families 17
    4.1.3 Affected members of a dyslexia family 18
    4.1.4 A Bull Terrier tail-chasing case-control cohort 18
    4.1.5 Zebrafish embryos 18
  4.2 Next generation sequencing 19
    4.2.1 Pooling strategy 19
    4.2.2 Library preparation and sequencing 20
  4.3 Data analysis 20
    4.3.1 Alignment of sequencing reads 20
    4.3.2 Detection of genetic variations 21
    4.3.3 Evaluation of allele frequency estimates 22
    4.3.4 Annotation and filtering of variants 23
    4.3.5 Association analysis 24
    4.3.6 Identification of NTRs 24

    - 4.3.7 Gene expression 25
  - 4.4 Experimental validation 25
    - 4.4.1 Genotyping 25
    - 4.4.2 PCR 26
    - 4.4.3 RT-PCR 26
    - 4.4.4 qRT-PCR 26
    - 4.4.5 Sanger sequencing 26
    - 4.4.6 Functional studies 26
- 5 Results 27
  - 5.1 Evaluation of pooling strategy (Paper I) 27
  - 5.2 Pooled exome sequencing and candidate gene for studying IS (Paper II) 28
  - 5.3 Whole genome sequencing of dyslexia susceptibility haplotype (Paper III) 28
  - 5.4 Identification of NTRs using RNA-seq (Paper IV) 29
- 6 Discussion 31
- 7 Conclusion and future perspective 33
- 8 Acknowledgements 35
- 9 References 38

LIST OF ABBREVIATIONS

AAF	Alternative allele frequency
AKAP2	A-kinase anchoring protein 2
APP	Amyloid beta precursor protein
BAM	Binary alignment map
bp	base pair
CAGE	Cap analysis of gene expression
CEP63	Centrosomal protein 63
CNV	Copy number variation
DCDC2	Doublecortin domain containing 2
DD	Developmental dyslexia
DNA	Deoxyribonucleic acid
dyx1c1	Dyslexia susceptibility 1 candidate 1
EMSA	Electrophoretic mobility shift assay
FANTOM	Functional ANnoTation Of the Mammalian genome
FBN1	Fibrillin 1
FBN2	Fibrillin 2
GATK	Genome Analysis Toolkit
GWAS	Genome-wide association study
HapMap	Haplotype map
HGP	Human genome project
HSPG2	Heparan sulfate proteoglycan 2
INDEL	Insertion or deletion
IS	Idiopathic scoliosis
kb	Kilo base pairs
LBX1	Ladybird homeobox homolog 1
LHX2	LIM homeobox 2
MAF	Minor allele frequency
Mb	Mega base pairs
NGS	Next generation sequencing
NTR	Novel transcribed region

OMIM	Online Mendelian Inheritance in Man
ORF	Open reading frame
PCR	Polymerase chain reaction
PE	Preeclampsia
POC5	POC5 centriolar protein
PSEN1	Presenilin 1
PSEN2	Presenilin 2
ptk7	Protein tyrosine kinase 7b
qRT-PCR	Quantitative reverse transcription polymerase chain reaction
ROBO1	Roundabout guidance receptor 1
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RT-PCR	Reverse transcription polymerase chain reaction
SAM	Sequencing alignment map
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SMRT	Small-molecule real-time
SV	Structural variation
S100B	S100 calcium binding protein B
TFBS	Transcription factor binding site
WGS	Whole genome sequencing

1 BACKGROUND

“Like father, like son” is probably the best-fitting proverb to describe heredity. Nowadays, almost everyone knows the causation of such patterns is inheritance. What are the factors in determining the variation of inherited characteristics? The first recorded successful attempt to answer the question is from Gregor Mendel, who discovered the law of inheritance and the heredity units, now called “genes”. Genes, lying on the double helix of DNA strands, are composed of four basic elements -- two purines, adenine (A) and guanine (G), and two pyrimidines, thymine (T) and cytosine (C). The combination and order of the four nucleotides cause the diversity of characters and traits in human beings. After the launch of the Human Genome Project (HGP) and release of the human genome sequence (Consortium, 2004; Lander et al., 2001; Venter et al., 2001), we have a clearer vision of the genes encoded in the human genome. At the same time, powerful bioinformatics software have been developed to decode the hidden messages written in our genome. The complete reference sequence and new tools allow us to dig more deeply into genomes for identifying the genes corresponding to or associated with specific phenotypes, especially disorders.

1.1 GENETIC VARIATIONS

The completion of the HGP revealed the sequences of 3 billion base pairs DNA, and more than 25,000 protein coding genes. Over 99% of DNA sequences in the human genome are consistent, while thousands of genetic variations have been observed across the human genome. We can divide them into different types based on their sizes: small-scale sequencing variation and large-scale structural variation.

1.1.1 Different types of genetic variations

The largest group of small-scale sequencing variations is the single nucleotide variant (SNV) (Figure 1a). An SNV is described as a substitution of a single nucleotide. DNA substitutions are classified as transitions (between two purines or two pyrimidines) and transversions (interchanges of purines for pyrimidines). The majority of the single base pair variations can have two forms of alleles, either transition or transversion. However, some polymorphic sites may be complex, e.g. three to four alternative alleles. SNVs with frequency in population of over 1% were termed as single nucleotide polymorphisms (SNPs). A total number of 1.42 million SNPs were identified in the human genome draft (Sachidanandam et al., 2001). In addition to SNVs, short insertions or deletions (INDELs) and short repeats distribute across all or part of the genome (Figure 1b). An INDEL is a deleted base(s) or extra inserted base(s) ranging from 1 bp to 1 kb in DNA sequencing. Tandem repeats are one kind of short repeats and can be divided into three categories according to the length of repeat units: satellites (over 50 bp), minisatellites (10-50 bp) and microsatellites (less than 10 bp). Most microsatellites have repeats of one to four nucleotides.

Large-scale structural variations (SVs) can be defined as the rearrangement of segments over 1kb (Feuk et al., 2006). Some major structurally abnormal variants include copy number variations (CNVs), inversions and translocations (Figure 1b). CNVs represent deletions, insertions or duplications of long tandemly repeated sequences. Although SNPs are most abundant in the genome, structural variations actually affect more numbers of nucleotides.

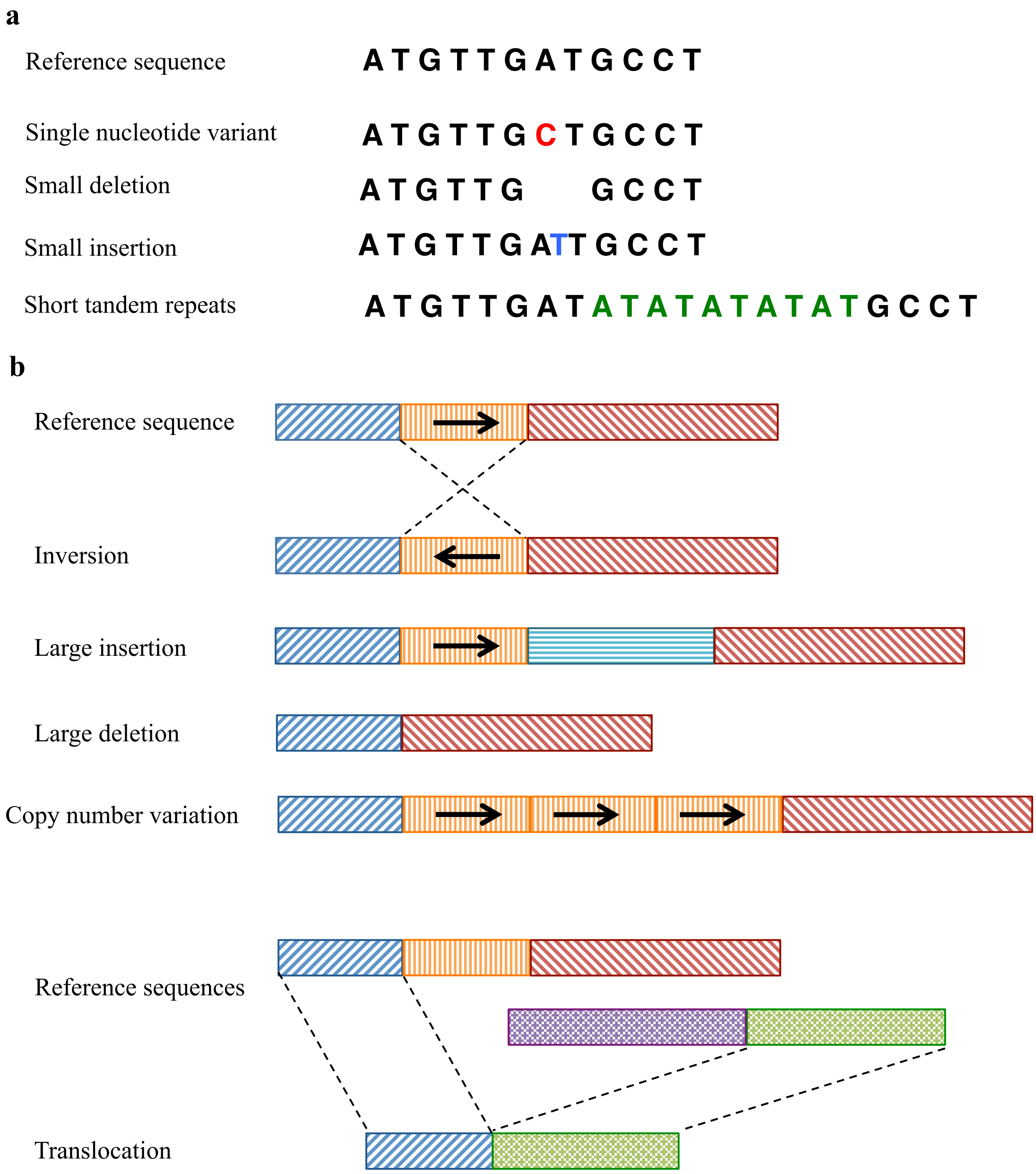
a
Reference sequence
ATGTTGATGCCT
Single nucleotide variant
ATGTTGCTGCCT
Small deletion
ATGTTG GCCT
Small insertion
ATGTTGATTGCCT
Short tandem repeats
ATGTTGATATATATATATATGCCT
b
Reference sequence
Inversion
Large insertion
Large deletion
Copy number variation
Reference sequences
Translocation

Figure 1. Different types of genetic variations a) Small-scale sequencing variation, b) Large-scale structural variation. Each color block represents a long DNA sequencing over 1kb.

1.1.2 Effects

A single nucleotide substitution in coding regions can change the encoded protein sequence, such as a missense variant. A nonsense mutation generates a stop codon and truncates the protein. A silent variant usually alters DNA sequences but not protein sequences. The point variants that can change the protein sequence are classified as non-synonymous variants, while those that do not affect protein sequence are synonymous variants. Any coding INDEL with a size at a non-integral number of codon causes a shift of open reading frame (ORF), consequently altering the protein sequence. Alternative splicing of primary transcripts can be affected by both SNVs and INDELs located near the splicing site. A large-scale structure variation will lead to loss dysfunction or duplication of the genes harbored in the SV regions. The genetic variations at non-coding regions may change the binding of transcription factors or regulation elements and DNA methylation, thus affecting the gene expression nearby.

There are more than 3 million genetic variants in two published individual diploid genomes (Levy et al., 2007; Wheeler et al., 2008). According to evolution, most of the variants in the human genome are harmless and do not have known effect on phenotypic traits. Certain genetic variations endow individual person with specific characteristics, e.g. skin color and eye color, body height and blood type. However, others may produce loss or gain of functions and result in disorder or susceptibility to diseases. For example, sickle cell disease and cystic fibrosis are determined by genetic variants in a single gene.

1.2 DISORDERS

1.2.1 Monogenic disorders

A disorder governed by a single gene with a clear pattern of Mendelian inheritance is called a monogenic (Mendelian) disorder. The two diseases mentioned in 1.1.2 are typical examples of monogenic disorders. There are five basic modes of Mendelian inheritance patterns: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive and Y-linked. In recessive mode, the phenotype is only expressed when the underlying locus is homozygous. In dominant mode, the phenotype is expressed as long as the underlying locus is heterozygous.

1.2.2 Complex disorders

The inheritance of most common disorders does not always follow Mendelian patterns. Disorders may be governed by multiple genetic loci and environmental factors, therefore defined as complex or multifactorial disorders. For example, most early on-set Alzheimer patients are caused by any mutation in genes *PSEN1*, *PSEN2* or *APP* in a autosomal dominant inheritance pattern (Bateman et al., 2012), while sporadic cases are affected by the environment as well (Gatz M et al., 2006). Complex disorders are usually determined by several gene loci with small effect (polygenic inheritance) or one locus with a major effect on the phenotype, modified by other genes with minor effects (oligogenic inheritance). Two complex disorders involved in the studies in this thesis are described below.

1.2.2.1 Idiopathic scoliosis

Scoliosis is a three-dimensional deformity of the spine as a lateral curvature of the spine in the coronal plane of more than 10° (Terminology Committee of the Scoliosis Research Society, 1976). It is usually classified into three major types: congenital (spinal abnormity at birth), syndromic (typically related neuromuscular, skeletal or connective tissue disorders) and idiopathic (causes unknown) (Altaf et al., 2013).

Idiopathic scoliosis (IS) constitutes the largest subgroup of human spinal curvatures (Gorman et al., 2012) and affects approximately 2-3% of the population worldwide (Luk et al., 2010; Rogala et al., 1978; Willner and Udén, 1982). The onset age can range from infant to adolescent, especially developing in childhood and adolescence. In a school-screen program, if there is a rib hump visible during an Adam's forward bend test, the individual will be sent for an X-ray. The clinical diagnosis of scoliosis is confirmed by a standing spinal radiograph showing a lateral curvature of the spine over 10 degrees. Even though the etiology of IS is poorly understood, heredity is recognized as an important contribution factor to it (Grauers et al., 2012; Wynne-Davies, 1968).

1.2.2.2 Developmental dyslexia

Dyslexia refers to a difficulty in reading and writing, in despite of normal senses, intelligence, and adequate education. It can be classified into two types: acquired dyslexia (caused by brain damage) and developmental dyslexia (DD). The typical symptoms of DD include spelling errors, slow naming of letters and words and poor verbal short-term memory. They can be mostly covered by phonological deficits (Ramus and Szenkovits, 2008). According to the epidemiological reports in the 1990s, the prevalence rates of DD range from 5 to 10% and possibly even to 17.5% (Shaywitz, 1998). It is the most common learning disability in childhood and may be diagnosed at school age after parents or teachers notice difficulties in reading and writing. The diagnosis test for dyslexia measures a series of reading-related cognitive skills, involving phonological decoding, phonological awareness, orthographic coding, rapid automatic naming, word recognition and spelling (Francks et al., 2002).

In addition to gender, environmental factors and co-occurrences with many other neurodevelopmental disorders (Friend et al., 2008; Gilger et al., 1992; Katusic et al., 2001; Paracchini, 2011; Rutter M et al., 2004), genetic factors are recognized as playing an important role in the etiology of dyslexia. Several twin studies demonstrate almost twice the concordance in monozygotic twins than that in dizygotic twins, suggesting a significant genetic component in dyslexia (Bakwin, 1973; DeFries et al., 1987; Harlaar et al., 2005; Hawke et al., 2006; Stevenson et al., 1987). In general, dyslexia shows complex genetic patterns: incomplete penetrance (Fagerheim et al., 1999), phenocopy (Nopola-Hemmi et al., 2001), heterogeneity of susceptibility loci in different families and polygenic influence (Fisher and DeFries, 2002).

1.3 GENE MAPPING IN DISORDERS

The crossover of homologous chromosomes usually occurs during prophase of meiosis I, causing recombination. In principle, the loci physically nearby are less likely to be segregated during recombination. Gene mapping aims to identify the underlying determinants of phenotypes based on co-segregation pattern of all loci, which are also called genetic markers. Some genetic variations, e.g. microsatellite and SNP (described in 1.1.1), are used as markers for gene mapping.

1.3.1 Linkage analysis

Linkage analysis focuses on the likelihood of recombination fraction between two loci in a family. The achievement of such analysis requires three parameters: a clear pattern of inheritance, highly polymorphic marker and penetrance of each genotype. Such test is called parametric analysis. In the late 20th century, microsatellites were employed as genetic markers in linkage studies. They were replaced by SNP arrays that can test much denser genetic markers along the genome simultaneously. Parametric analysis succeeds in mapping disease locus in monogenic disorders that show clear Mendelian inheritance patterns. However, it is very difficult to map susceptibility loci to complex disorders because the genotype and phenotype usually are in incomplete correspondence (Altmüller et al., 2001). To overcome this problem, a different approach is to seek the chromosomal segments shared by affected members in the family regardless of inheritance pattern. Since the genetic mode parameter is missing in the analysis, this approach is called non-parametric analysis.

In the case of DD, most cases show a pattern of complex traits. However, some large or multiple small families displayed Mendelian inheritance patterns. Accordingly a series of genetic linkage studies have been performed (Kere, 2011). By far, nine gene loci were pinpointed and considered as dyslexia susceptibilities in the OMIM database (http://omim.org/). In a linkage study of a three-generation Finnish dyslexic family, a 33 Mb region (3p12-q13) on chromosome 3 was found to be associated with dyslexia (Nopola-Hemmi et al., 2001). The same region (3q13) was supported by a genome-wide scan on dyslexic families in the US (Fisher et al., 2002). The susceptibility loci on chromosome 3 were later reported to be linked to speech-sound disorder (Stein et al., 2004). Follow-up studies investigating the linkage region indicated *ROBO1* as a susceptibility gene for dyslexia (Hannula-Jouppi et al., 2005; Lamminmäki et al., 2012).

1.3.2 Association studies

Another approach to reveal the incomplete relationship between genotype and phenotype is association study. The rationale of association study is based on linkage disequilibrium (LD), which is a non-random association of allele at two or more loci. As a result of linkage analysis on complex disorders, the susceptibility loci may be several Mb, which normally harbor many genes. For this reason, extra work and effort on studying the candidate genes in the loci need to be addressed. The association study is usually performed on a group of unrelated affected individuals and a group of unrelated controls. If there is an association between a phenotype and genotype, the probability of seeing the specific genotype will be greater than by chance in an individual with the phenotype.

1.3.2.1 Genome-wide association studies

In 2002, the International HapMap Project was launched, aiming to construct a map of common genetic variants in the human genome (Gibbs et al., 2003). The project investigates the genotypes of sequencing variants, their frequencies and the link between them within different populations. It provides vast tag SNPs to capture large proportion of variants for association study on genome-wide scale. At the same time, the high-throughput microarray enables the investigation of thousands of SNPs simultaneously at an inexpensive cost. A genome-wide association study (GWAS) genotypes a dense set of SNPs across the genome to investigate common variants for their links to disorders and quantitative traits (Hirschhorn and Daly, 2005). Compared with genome-wide linkage scan, a GWAS is more powerful for identifying disease-associated variants with moderate effect (Risch and Merikangas, 1996). Consequently, it is an efficient and powerful tool for analyzing the genetic architecture of complex disorders. When testing multiple SNPs on a genome-wide scale, each SNP represents one hypothesis. Accordingly, the significance levels should be corrected for the number of SNPs. Assuming a power of 0.5, the threshold of genome-wide significance is 5×10^{-7}.

The first published GWAS was performed on a cohort of about 14,000 cases and 3,000 shared controls (Burton et al., 2007). This study discovered several genome-wide associations ($P < 5x10^{-7}$) between genotypes and human complex disorders, e.g. coronary artery disease, Crohn's disease, rheumatoid arthritis, bipolar disorder, type I and type II diabetes. The GWAS catalog (https://www.ebi.ac.uk/gwas/) was initially developed for collecting SNP-trait associations reported from published genome-wide association studies (Hindorff et al., 2009). For the past 10 years, more than 10,000 SNPs were identified to be strongly associated ($P< 5x10^{-8}$) with complex traits (Welter et al., 2014).

In the case of IS, a GWAS of a Japanese case-control cohort identified a SNP (rs11190870) near *LBX1* significantly associated with IS (Takahashi et al., 2011). The same case and control sample sets demonstrated more association The strong association of *LBX1* with IS was replicated in Chinese and Caucasian populations (Fan et al., 2012; Gao et al., 2013; Jiang et al., 2012; Londono et al., 2014). Several other gene regions have been reported to show association with IS in Asian and Caucasian population (Kou et al., 2013; Miyake et al., 2013; Sharma et al., 2011), however they are not easy to replicate in other populations.

1.4 MISSING HERITABILITY OF COMPLEX DISORDERS

Even though linkage study and genome-wide association study have demonstrated great success in mapping disease-associated genes, the identification of genetic variations that contribute to complex disorders was becoming slow and arduous (Hardy and Singleton, 2009). Linkage analysis provided amounts of knowledge on high-risk rare alleles underlying monogenic disorder. Though those variants have high penetrance, their frequencies in the population are extremely rare. Beyond linkage study, GWAS is based on the hypothesis that common disease may be caused by common variants with high to low effects (Collins et al., 1997). However, the screening for common variants resulted in lower impact of genetic factors on complex traits than we expected. For example, the heritability of age-related macular degeneration may be 50% explained by small amount of common variants that have high effects (Maller et al., 2006). On the other hand, more than 40 genetic loci can only explain about 5% of the heritability of height (Visscher, 2008). According to the minor allele frequency (MAF) in the tested population, genetic variations are categorized into rare (MAF

< 1%), low-frequency (MAF between 1% to 5%) and common variants (MAF > 5%). If we look at Figure 2, monogenic disorder causative rare variants and common variants with high or low impact lying on the corners of the spectrum can be detected by family-based linkage study or GWAS. There is a large portion of low-frequency variants with moderate effect that cannot be covered by traditional genetic approaches.

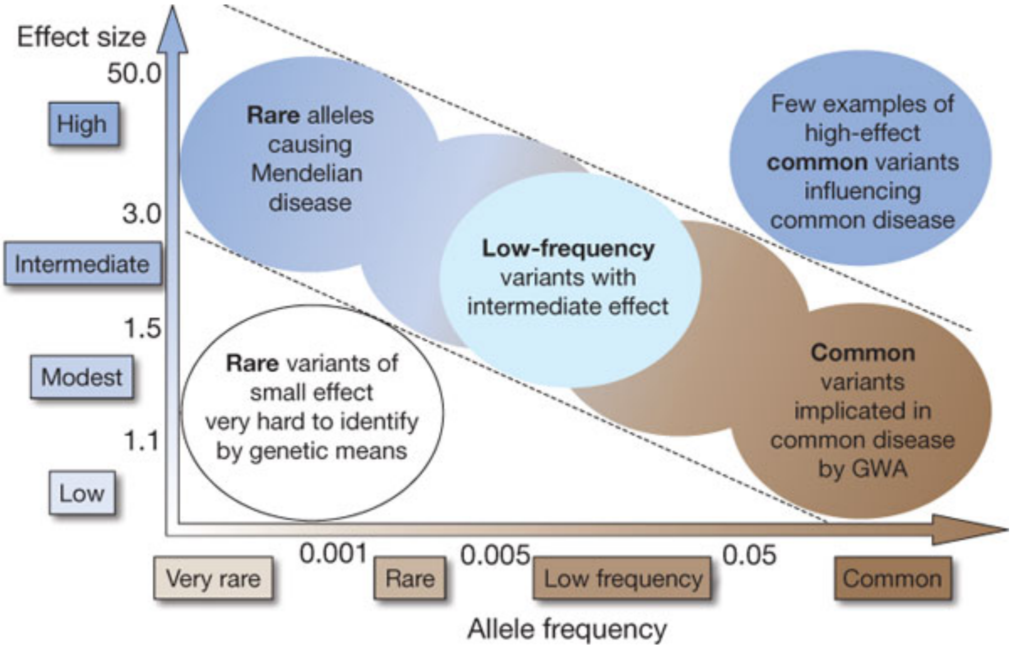
Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 2. Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from (McCarthy et al., 2008). The reprint permission of this figure from (Manolio et al., 2009) was granted by Nature Publishing Group.

Mapping of susceptibility gene in complex disorder is far from a simple task. Several explanations have been suggested for explaining missing heritability (Manolio et al., 2009). Firstly, there may be more common variants with low impact on complex disorders. Secondly, aside from common variants, rare and low-frequency variants likely contribute to phenotype with large effects that are poorly detected. Thirdly, the structure variations that affect larger regions than point mutations are poorly captured by SNP arrays. In addition, low power to detect gene-gene interaction and inadequate estimation for shared environment are other potential sources of missing heritability. For these reasons, there is a large demand for developing a new technique to investigate rare and low-frequency variants, as well as structure variation.

1.5 FUNCTIONAL STUDIES FOLLOWING IDENTIFICATION OF DISEASE-CAUSATIVE GENES

To understand the effect of genetic variants and susceptibility genes on a disorder, many approaches at different levels have been applied. The computational predictions based on experimental data are usually the first step to study the function of genetic variants. In addition to computational approaches, experimental investigation is more than required for understanding gene function. Manipulating cultured human cell lines is a standard method in detecting the consequence of genetic variants and gene function *in vitro*. Furthermore, to mimic the gene regulations *in vivo*, animal models are widely used for studying human disorders.

1.5.1 *In silico*

The function of genetic variation can be predicted according to the genomic location. Missense, nonsense, splicing mutation and frame-shift INDEL directly change the protein sequence. Several tools have been developed for predicting the effect of amino acid substitution on protein structure and function based on machine, e.g. SIFT, PolyPhen-2 (Adzhubei et al., 2010; Ng and Henikoff, 2001; Ramensky et al., 2002), based on either the conservation of homolog protein sequence or the physical property of amino acid. To investigate variants in non-coding regions that may be involved with regulation of gene expression, the information from ENCODE (Encyclopedia of DNA Elements) project can be employed. ENCODE provides a variety of data generated with different assays to understand functional elements in the human genome, including DNA methylation, transcription factor binding site (TFBS) and RNA binding site, etc. (Consortium, 2012b).

1.5.2 *In vivo* – model organisms

Even though the functional study in cells has provided insight into gene regulation, it cannot mimic the interactions between different types of cells in a living organism. Moreover, certain types of cell line are very difficult to culture. Therefore, it is necessary to study gene function and simulate the dysfunction of candidate genes in model organisms. One approach to perform genetic manipulation in model organisms is to make a transgenic animal by introducing an exogenous DNA with a genetic mutation into zygotes. Another way is to modify the target sequence in an endogenous gene with a mutant version. To investigate the consequence of deletion and translocation, a candidate gene is knocked out by site-specific recombination, RNAi pathway or blocking their expression by using morpholino antisense oligonucleotides.

Because of its short generation time and large amount of eggs, zebrafish has become a principle vertebrate model for characterizing the function of genes involved in development and disease, especially when the sample of functional tissue is not available to access in a living human. To understand the function of certain susceptibility genes in dyslexia and IS, several studies have already been performed on zebrafish. The dyslexia candidate gene orthologue in zebrafish *dyx1c1* proved to control cilia growth, axonemal dynein assembly and cilia motility (Chandrasekar et al., 2013; Tarkar et al., 2013). The zebrafish *ptk7* mutants developed distinct spinal deformities, and later *ptk7* appeared to be involved in cilia mobility that is required for spine morphogenesis (Grimes et al., 2016; Hayes et al., 2014). Three functional variants on human IS-associated *POC5* were examined on the zebrafish model and the expression of those variant RNAs demonstrated spine deformity (Patten et al., 2015). Moreover, the overexpression of *lbx1* homologs in zebrafish causes body axis deformation, including defective convergent extension movement and body curvature (Guo et al., 2016).

2 INTRODUCTION

A technique developed for sequencing DNA enabled the reading of the nucleotide sequence of a whole gene in the 1970s (Sanger et al., 1977). This technique was applied in the HGP for sequencing the human genome in the 1990s, and it took more than 10 years to complete the HGP. The Sanger-based method was regarded as the first generation of DNA sequencing. In the last decade a new technique that can produce millions or billions of DNA sequence fragments in a short time has been innovated to replace the slow and laborious first generation sequencing with limited output at a high cost. Accordingly the new method is referred to as next generation sequencing (NGS).

2.1 NGS PLATFORMS

The procedure of NGS is composed of two major parts: 1) DNA template preparation and 2) sequencing and imaging. The first step for preparing template is fragmentation. DNA templates were randomly sheared into small-size fragments. For short-read NGS, the length of fragment ranges from 50 bp to 400 bp, depending on platform and library kit, while the long reads could be up to 100 or 200 kb in length for long-read NGS (Table 1). The template fragments were then ligated to adaptors for amplification (except SMRT method) and sequencing. Illumina, SOLiD and Complete Genomics platforms are employed for genetic and genomic studies in this thesis, so the sequencing approaches are briefly described below.

Table 1. Brief summary of NGS platform, adapted from (Goodwin et al., 2016)

Sequencing methods	Template preparation	Platforms	Read length
Sequencing by ligation	Emulsion PCR	SOLiD (Thermo Fisher)	50-75 bp
	In-solution DNA nanoball generation	Complete Genomics (BGI)	50-100 bp
Sequencing by synthesis (cyclic reversible termination)	Solid-phase bridge amplification	Illumina	36-250bp
	Emulsion PCR	Qiagen	NA
Sequencing by synthesis (single-nucleotide addition)	Emulsion PCR	454 pyosequencing (Roche)	Up to 1 Kb
	Emulsion PCR	Ion Torrent (Thermo Fisher)	200-400 bp
Single-molecule real-time (SMRT) sequencing	Single molecule	PacBio	8-12 kb
	Single molecule	Nanopore	Up to 200 kb
Synthetic long-reads	Solid-phase bridge amplification	Illumina	~100 kb
	Emulsion PCR	10X Genomics	Up to 100 kb

2.1.1 Illumina

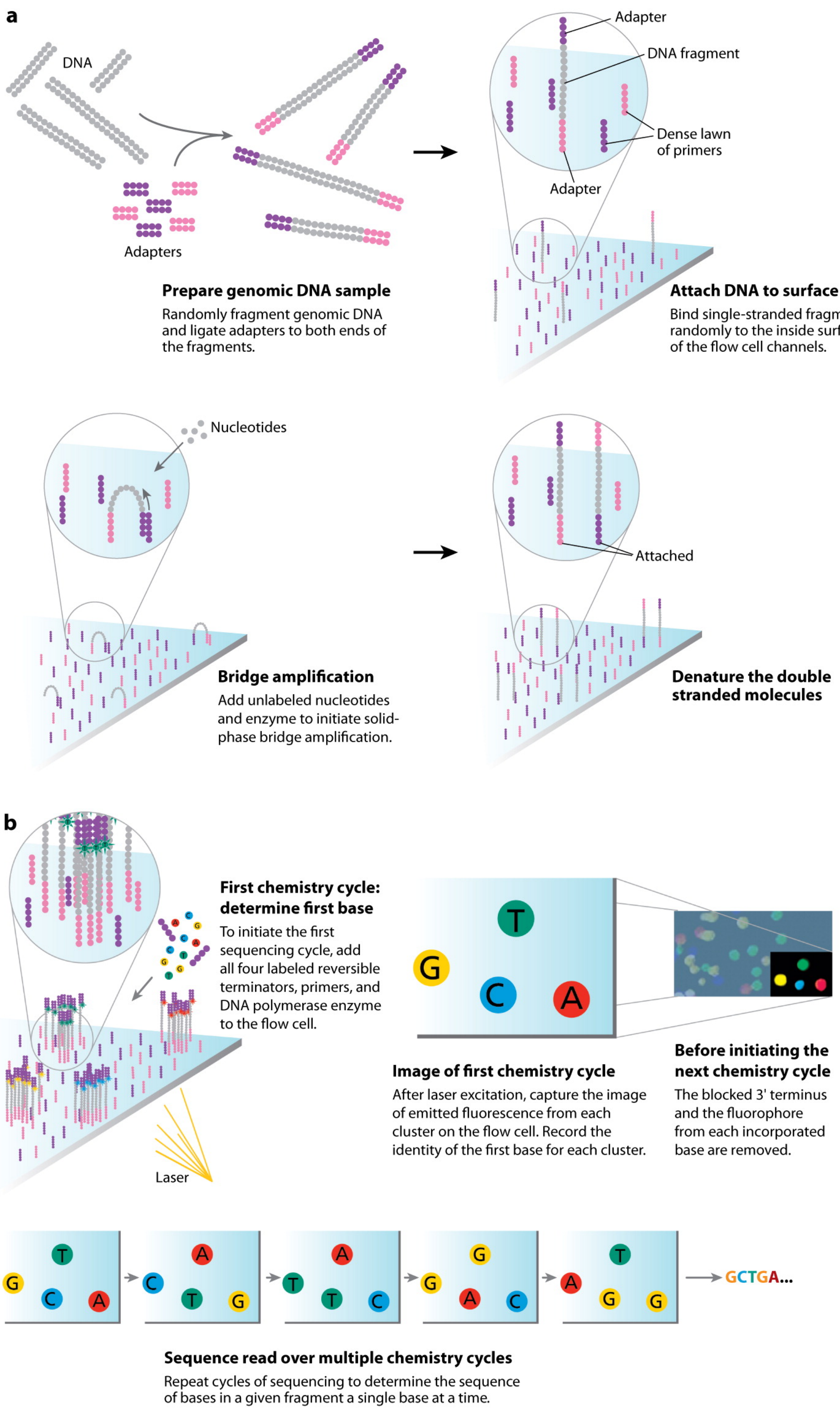
a
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Nucleotides
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Attached
Denature the double stranded molecules
b
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.
T
G
C
A
Laser
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.
GCTGA...
Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Figure 3. The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3’-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3’-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation. (Mardis, 2008)

In the Illumina platform, the adapted DNA templates are randomly ligated on the solid slide surface of a flow cell. The inside surface contains both forward and reversed primers complementary to the adaptors of the DNA template. Then the bridge amplification is then performed to enrich the spatially separated DNA template clusters (Figure 3a). After detaching one end of primer, the DNA clusters are sequenced by cyclic reversible termination approach. Four 3'-blocked deoxynucleotides (dNTPs) labeled with fluorescent and DNA polymerase are simultaneously added to the flow cell channel. After the incorporation of a dNTP to a template cluster, the emitted fluorescent is captured to determine the identity of the incorporated nucleotide of each cluster in the current cycle (Figure 3b). Before moving to next cycle, a cleavage step is performed to remove the terminus and the fluorescent label, together with an additional wash.

2.1.2 SOLiD

In the SOLiD platform, the DNA templates attached with adaptors are ligated to magnetic beads that are located on the glass slide. The emulsion PCR is performed to enrich the template sequence on beads. The sequencing by ligation approach used in SOLiD involves hybridization and ligation of primed template and fluorescently labeled anchor-probe with known base(s). DNA ligase is added to join the labeled probe to the primer where the probe was hybridized to the template. Those free probes are then washed away. In each cycle, the emitted fluorescence of ligated probe is captured for imaging the template. Unlike other sequencing platforms, SOLiD utilizes two-base-encoding probes, in which each ligation signal represents two bases of nucleotides, defined as color-space. After removing the probe-anchor complex, a new round of hybridization and ligation cycle starts again.

2.1.3 Complete Genomics

Complete Genomics is an approach performing template enrichment in solution, through the process of rolling cycling amplification. In this case, DNA sample is processed to create a circular template with four distinct adaptor regions, and then amplified to generate DNA nanoballs that are immobilized on a flow cell (Drmanac et al., 2010). The sequencing procedure is similar to the SOLiD platform, but uses one-base-encoding probes, so each fluorescent signal can directly represent one single nucleotide.

2.2 NGS APPLICATION

In recent years, high-throughput NGS technology has revolutionized genetic and genomic studies by providing the opportunity for interrogating different properties of cellular processes at massive scale. It covers genome re-sequencing, whole genome and transcriptome *de novo* assembly, transriptome profiling, DNA-protein interaction, and epigenetic characterization. Below are three types of NGS applied in this thesis.

2.2.1 Whole genome sequencing and exome sequencing

Whole genome sequencing (WGS) is to determine the complete DNA sequencing of a whole genome of any organism at one time. Using this method, an entire human genome can be resequenced within a single day, instead of a decade spent in the HGP with Sanger-based sequencing. The cost of sequencing per base pair nucleotide is dramatically reduced as well (Metzker, 2010; Shendure and Ji, 2008). In addition to the benefits associated with the time and costs of sequencing whole genome, the new method also facilitates the discovery of novel genetic variations that cannot be accomplished by DNA arrays with pre-defined probes. High though-put exome sequencing was designed to re-sequence the entire exome in a human genome (Mamanova et al., 2010). It needs an extra step in DNA library preparation to capture the exons of annotated genes. Exome sequencing covers roughly 1% of the entire human genome (Hodges et al., 2007), and enable the target re-sequencing on more samples in a sequencing run with high coverage and depth at a limited cost.

2.2.1.1 Population

A major advantage of high though-put WGS and exome sequencing is to identify new genetic variations. In the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/), an increased trend in the total numbers of variant records has been observed over the past ten years (Figure 4).

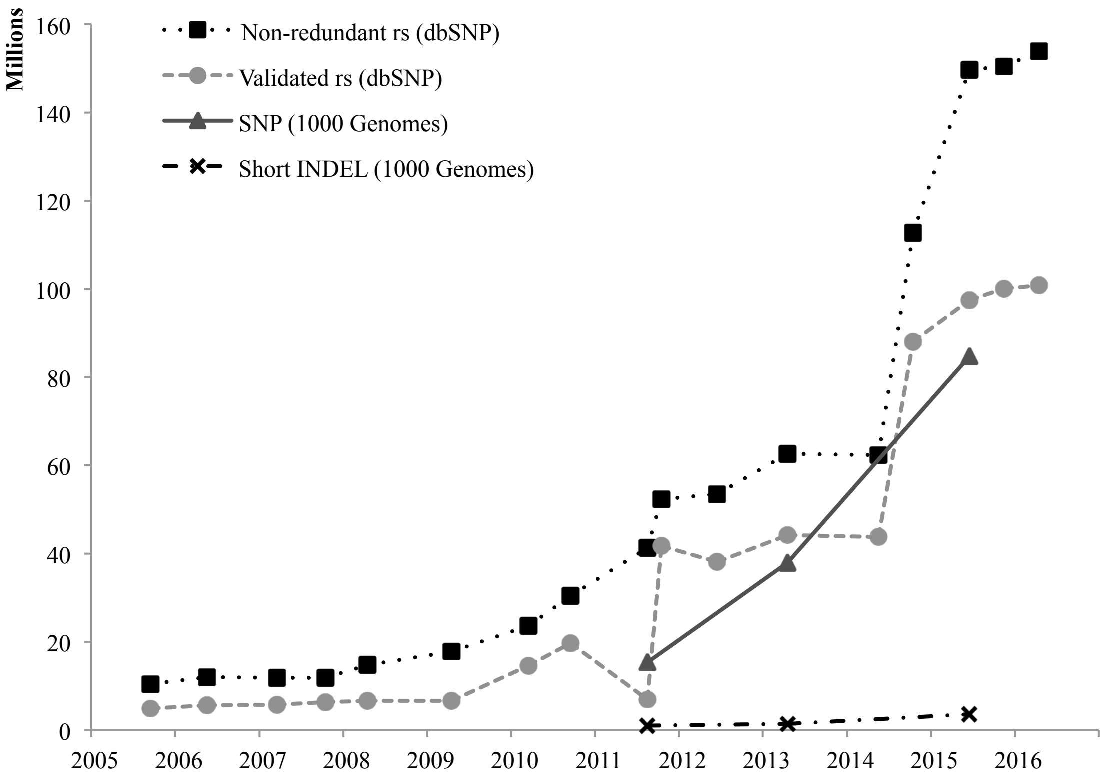

Figure 4. Rate of growth in dbSNP database and 1000 Genomes project (Human) The statistic summary was retrieved from the dbSNP database and publications of 1000 Genomes project (Consortium, 2010, 2012a; The 1000 Genomes Project Consortium, 2015).

In 2010, the 1000 Genomes project was released, aiming to provide a characterized map of human genetic variations. From the WGS of 179 individuals in four populations, the exome sequencing of 697 individuals in seven populations and the deep sequencing of two family trios, 15 million SNPs, 1 million INDELs, and 20,000 structural variations were discovered (Consortium, 2010). The genome locations, allele frequencies in populations and haplotypes were described about those variations, most of which were not annotated previously. After five years, the project has dramatically expanded to 2504 individuals from 26 populations, more than half of which have been performed whole genome and exome sequencing. The total numbers of genetic variants rose to 88 million (Figure 4) (Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2015). More sequencing with large sample sets on population levels have been performed in the UK and Iceland (Gudbjartsson et al., 2015; The UK10K Consortium, 2015), as well as in Finland (http://www.sisuproject.fi). The insight into population genetic properties of variants provided references and facilitated identifying variants associated with human disease in a specific population. In our case, the references of allele genotype and frequency in European and Finnish populations played important roles in filtering candidate variants for validation.

2.2.1.2 Disorders

Both WGS and exome sequencing have successfully identified disease-associated variants, especially in Mendelian disorders (Bamshad et al., 2011; Ng et al., 2010). One of the earlier publications using exome sequencing for identifications of disease causative genes was reported on a study of Freeman-Sheldon syndrome (Ng et al., 2009). High-thoughput DNA sequencing was also utilized for clinical or molecular diagnosis (Choi et al., 2009; Lupski et al., 2010). In trio studies, the exome sequencing method also helped identify *de novo* mutations associated with complex disorders (Iossifov et al., 2014; O'Roak et al., 2011; Sanders et al., 2012; Vissers et al., 2010). There were also several NGS studies of IS and DD. Exome sequencing has been performed on individual samples in families or in cohorts. Some rare coding variants were discovered in *AKAP2*, *FBN1*, *FBN2*, *HSPG2*, *POC5* and six musculoskeletal collagen genes showing genome-wide significance associate with IS, (Baschal et al., 2015; Buchan et al., 2014; Haller et al., 2015; Li et al., 2016; Patten et al., 2015). A novel mutation in *CEP63* was identified in a Swedish family co-segregating with dyslexia by exome sequencing (Einarsdottir et al., 2015). A missense polymorphism in *DCDC2* and a non-coding variant in *S100B* were reported to be associated with dyslexia by using high-throughput sequencing on candidate genes (Matsson et al., 2015).

2.2.1.3 Pooled DNA sequencing

WGS and exome sequencing are becoming increasingly popular in genetic studies because of their wide coverage and single-base resolution. However, in association study, sequencing each individual in the population scale is still costly, laborious and time-consuming. An alternative approach is to pool a number of individual DNA samples and sequence the pooled DNA, which can efficiently increase sample size and sequencing depth at a reduced cost and effort in library preparation. This approach only uses one library for entire pooled samples instead of separate libraries as sequencing individuals. Pooled DNA sequencing is designed for different purpose, such as genotype-phenotype mapping, domestication, genome evolution and cancer study (Schlötterer et al., 2014). The key for genotype and phenotype mapping is to compare the allele frequency between groups of individuals. If the individuals are divided into pools according to their phenotype, the phenotype-linked variants will show large differences in frequency between pools (Sham et al., 2002). Meanwhile, the essence of pooling strategy is the allele frequency estimates in a group of individuals. Therefore, pooled DNA sequencing is appropriate for detecting phenotype-associated variants at a large scale and low cost. It can be applied in WGS, exome sequencing and target sequencing on restriction sites.

The pooling strategy showed good performance in fine mapping complex traits in model organisms (Axelsson et al., 2013; Bastide et al., 2013). Furthermore, pooled DNA sequencing demonstrated excellent performance in detecting rare variants when exploring candidate genes in human samples as well (Bansal et al., 2010; Druley et al., 2009; Out et al., 2009). In addition, certain pooled DNA sequencing studies at the whole exome scale have also discovered low-frequency variants associated with human complex disorders (Flanagan et al., 2013; Jiao et al., 2014).

2.2.2 RNA sequencing

RNA sequencing, so-called “RNA-seq”, is a method to study mRNA expression levels in a transcriptome-wide range using deep sequencing technology. The library preparation in RNA-seq requires converting a population of total RNA or polyA-tailed RNA to a library of cDNA fragments. With or without amplification, the cDNA fragments were then sequenced in a high-throughput approach. Before the era of NGS, large-scale gene expression studies were performed using microarrays, which has limitations in accuracy of low-abundant gene expression measurement and background level of hybridization (Marioni et al., 2008; Wang et al., 2009). This technique also requires the existing knowledge of genome annotation for profiling the whole transcriptome. Unlike microarrays, RNA-seq does not rely on the pre-determined probe and is thus able to detect transcription start sites, capture alternative splicing patterns and construct novel transcripts (Ozsolak and Milos, 2011; Wang et al., 2009).

RNA-seq has been applied to profiling mouse and human transcriptomes (Mortazavi et al., 2008; Sultan et al., 2008), especially during early developmental stages at single-cell resolution (Töhönen et al., 2015; Xue et al., 2013; Yan et al., 2013). In addition to the profiling of mammalian transcriptomes, RNA-seq has been utilized for transcriptome characterization and genome annotation on other model organisms (Derrien et al., 2011; Derti et al., 2012; Tan et al., 2013). As an invaluable model organism mentioned in section 1.5, zebrafish was used for understanding vertebrate development and human disease. To study gene expression at early development, several RNA-seq studies have been performed on zebrafish (Aanes et al., 2011, 2013; Vesterlund et al., 2011; Yang et al., 2013), from which novel transcripts and isoforms were discovered before and after zygotic genome activation. In addition, tag-based approach, cap analysis of gene expression (CAGE), was employed to map the transcription initiation and predict the transcription start sites in zebrafish embryo development (Nepal et al., 2013). RNA-seq has provided a large volume of data for identifying novel transcribed regions (NTRs) and isoforms of known genes. Before RNA-seq was used for annotating the zebrafish genome, there were 28717 transcripts in Ensembl genome annotation release 54 (reference genome Zv8) in early 2009. Until the end of 2014, the total number of transcripts was almost double (56754) in Ensembl genome annotation release 79 (reference genome Zv9). The new zebrafish genome annotation is still under construction based on reference genome GRCz10. The total number of transcripts has so far increased to 57866 (release 84).

2.3 BIOINFORMATIC ANALYSIS OF NGS DATA

In the Illumina platform, the raw reads generated by sequencing machine are displayed in FASTQ format, regarded as the standard layout format of sequencing reads. Except for the *de novo* assembly purpose, the raw reads are usually aligned to the reference genome. The alignment procedure is also called mapping. The color-space reads generated by SOLiD platform are in different formats. When performing mapping, the reference genome should be converted to color-space to be consistent with sequencing reads. The mapped reads are usually presented in SAM format and further compressed to BAM format (Li et al., 2009). The BAM file contains information about the mapping quality and position of short reads. PCR duplicate removal is usually conducted on mapped DNA sequencing reads. A best practice workflow recommends INDEL realignment and base quality score recalibration on mapped reads (DePristo et al., 2011). After the pre-processing steps, the mapped reads are utilized for further analysis with diversity of purpose, e.g. variant detection, gene expression profiling, and new transcripts identification.

WGS and exome sequencing has been widely used in genetic research for identifying novel causal variants associated with rare diseases and susceptibility genes to common diseases. The capacity of sequencing instruments is now becoming increasingly large. A huge volume of both raw and analyzed data have been generated for the past several years and continue to increase especially when the sequencing cost drops dramatically (Goodwin et al., 2016). How shall we store large datasets, to evaluate the accuracy and precision, and to prioritize candidate genes? Data maintenance, analysis and interpretation are big challenges for handling NGS applications.

3 AIMS

The study projects presented in the thesis focused on the application of NGS technology in dissecting the genetics of complex diseases and annotating the model organisms used for functional study.

Paper I	Assess the reliability of rare and low-frequency variant discovery and evaluate the accuracy of minor allele frequency estimates with pooled DNA sequencing
Paper II	Identify susceptibility variants to scoliosis by exome sequencing of IS patients with a severe phenotype Investigate the association of previously reported SNPs with IS in a Scandinavian cohort by a candidate gene study
Paper III	Find the genetic variants within previously identified susceptibility locus that could explain the dysfunction of ROBO1 by re-sequencing the whole genome
Paper IV	Discover NTRs in zebrafish by using RNA sequencing data and an in-house pipeline

4 MATERIALS AND METHODS

4.1 STUDY SUBJECTS AND MATERIALS

4.1.1 Idiopathic scoliosis (IS) case-control cohorts

Between 2004 and 2013, IS patients were invited to participate in the IS and Genetics in Scandinavia (ScoliGeneS) project (Andersen et al., 2006; Grauers et al., 2013, 2014). It is a multi-center study involving Karolinska University Hospital in Stockholm, Sahlgrenska University Hospital in Gothenburg, Sundsvall and Härnösand County Hospital, Umeå University Hospital and Middelfart Hospital in Denmark. Until the end of 2013, 1885 individuals were eligible and available for inclusion in the Scoliosis and Genetics in Scandinavia (ScoliGeneS) project and DNA samples were successfully extracted from 1739 of them (blood from 98% of the samples and saliva from 2% of the samples). The controls are composed of two cohorts: 909 elderly women from the OPRA cohort and 903 young women from the PEAK-25 cohort living in Malmö, Sweden (Gerdhem and Akesson, 2007; McGuigan et al., 2007). DNA samples were extracted from the blood of the controls.

In **Paper I and II**, 100 severe IS patients in the ScoliGeneS cohort (24 male and 76 female) were selected for exome sequencing with pooling strategy and validation by individual genotyping. In **Paper II**, 1739 cases and 1812 controls were used for studying the association of *LBX1* with IS by genotyping. The ethical permits for sample collection and genetic study were approved by the Lund university Research Ethics Committee, the Regional Ethical Board in Skåne, the Karolinska Institutet Research Ethics Committee, the Regional Ethical Board in Stockholm and the Regional Committees on Health Research Ethics for Southern Denmark.

4.1.2 Preeclampsia (PE) case-control cohorts and families

The Finnish Genetics of Pre-eclampsia Consortium (FINNPEC) case-control cohort (Kaartokallio et al., 2014) was collected from five Finnish university hospitals from 2008 to 2011: Turku University Central Hospital, Oulu University Hospital, Helsinki University Central Hospital, University of Helsinki and Kuopio University Hospital. In **Paper I**, 90 severe PE patients with early onset or high proteinuria from the FINNPEC cohort were selected for exome-sequencing. In addition, 10 PE patients recruited from Kainuu and Helsinki were also included in exome sequencing. They belong to separate families chosen from previous family studies (Laivuori et al., 2003; Majander et al., 2013). The genomic DNA was extracted from the whole blood of the 100 samples. In addition to exome sequencing, they were also used for validation by individual genotyping. The ethical permits for sample collection and genetic study were approved by the Karolinska Institutet Research Ethics Committee and the Coordinating Ethics Committee of The Hospital District of Helsinki and Uusimaa.

4.1.3 Affected members of a dyslexia family

A multiplex three-generation family segregating dyslexia in an autosomal dominant fashion was identified in Helsinki, Finland (Nopola-Hemmi et al., 2001, 2002). All dyslexic and normal individuals in the family were native Finnish speakers. Twenty-one pedigree members affected with dyslexia were included for genotyping. Linkage analysis found out 19 of them shared the same haplotype on chromosome 3. In **Paper III**, the 19 affected individuals were included for whole genome sequencing using pooling strategy. In addition, two of them were selected for whole gnome sequencing individually. The ethical permits for sample collection and genetic study were approved by the Karolinska Institutet Research Ethics Committee and the Coordinating Ethics Committee of The Hospital District of Helsinki and Uusimaa.

4.1.4 A Bull Terrier tail-chasing case-control cohort

In a previous study about environmental effects on behavior, DNA samples were collected from tail-chasing Bull Terriers and controls (Tiira et al., 2012). To determine the tail-chasing behavior, detailed questionnaires were developed for dog owners. In **Paper I**, the DNA samples extracted from the blood of 10 Bull Terriers with tail chasing (3 male, 7 female) and 10 Bull Terriers without such phenotype (3 male, 7 female) were selected for whole genome sequencing. The same materials were used for validation by individual genotyping. The ethical permits for DNA collection and genetic studies were approved by the Finnish National Animal Ethics Committee.

4.1.5 Zebrafish embryos

The RNA-seq was performed using zebrafish material described in previous publication (Vesterlund et al., 2011). In brief, the eggs were collected immediately after fertilization, further staged by developmental time and observations of developmental stage: 1-cell, 16-cell, 512-cells and 50% epiboly. In each embryonic stage, total RNA was extracted from approximately 150 embryos. In **Paper IV**, we re-use the RNA-seq data for novel transcript identification. For experimental validation of putative NTRs, zebrafish cDNA from 50% epiboly was used as the template. Three biological replicates of each developmental stage were used for validating and measuring the expressions of randomly selected NTRs. The ethical permits for material collection and gene expression studies were approved by the Stockholm Animal Research Ethical Board.

In order to investigate the functions of candidate genes and the effects of genetic variants on non-coding regions, cell lines are usually employed to test the loss or gain of putative protein binding site on mutated loci. In **Paper III**, various types of cell lines were employed for functional study, which are describe in detail in the published paper.

4.2 NEXT GENERATION SEQUENCING

4.2.1 Pooling strategy

In **Paper I** and **II**, every 10 of a total of 100 IS patient samples were pooled together for exome sequencing. The same pooling scheme was utilized in **Paper I** for pooling the 100 severe PE patient samples and 20 Bull Terrier samples according to their characteristics (Table 2). In both exome sequencing studies of IS and PE, 800 ng of DNA extracted from individual samples was used for pooling. In the WGS study of Bull Terriers, 1 µg of DNA extracted from each Bull Terrier was used for pooling. The concentration and purity of DNA in the samples were controlled using a Nanodrop spectrophotometer, agarose gel electrophoresis and Qubit fluorometer. In **Paper III**, the same amount of DNA samples from 19 dyslexic family members were pooled together for whole genome sequencing.

Table 2. Pooling scheme for whole genome sequencing and exome sequencing

Study	Sequencing	Pool	#Samples	Characteristic
IS (Paper I, II)	2x100bp paired-end exome sequencing	1	10	Cobb angle: 54.9±10.3 from Skåne
		2	10	Cobb angle: 59.9±13 from Skåne
		3	10	Cobb angle: 54.9±9 from Skåne
		4	10	Cobb angle: 57.2±23.8 from Skåne
		5	10	Cobb angle: 54.7±8.2 from Skåne
		6	10	Cobb angle: 54.6±5.5 from Skåne
		7	10	Cobb angle: 59±19.5 from Stockholm
		8	10	Cobb angle: 55.6±9.4 from Stockholm
		9	10	Cobb angle: 56.3±17.4 from Stockholm
		10	10	Cobb angle: 55.9±9.3 from Skåne and Stockholm
PE (Paper I)	2x100bp paired-end exome sequencing	1	10	Severe and early onset PE
		2	10	Severe and early onset PE
		3	10	Severe PE with proteinuria over 6g/24h
		4	10	Severe PE with proteinuria over 6g/24h
		5	10	Severe PE with previous miscarriage(s)
		6	10	Severe PE with proteinuria over 3g/24h
		7	10	Severe PE
		8	10	Severe PE
		9	10	Severe and recurrent PE
		10	10	Familial PE
Bull Terrier (Paper I)	2x100bp paired-end WGS	1	10	Compulsive tail-chasing behavior
		2	10	Non tail-chasing
Dyslexia (Paper III)	2x100bp paired-end WGS	1	19	Dyslexic family members with shared haplotype

4.2.2 Library preparation and sequencing

For pooled exome sequencing in **Paper I** and **II**, the pooled genomic DNA samples were sheared to 300bp using a Covaris S2 instrument (Covaris, MA, USA). The libraries were prepared using SureSelectXT Reagent kits (Agilent Technologies, CA, USA) and an Agilent NGS workstation according to the manufacturer's instructions (SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing, version A). The SureSelectXT Human All Exon 50 Mb capture library was used for the targeted enrichment. Because of the high PCR duplication rate in the first round in the scoliosis DNA samples, technical replications were performed with the same library in eight pools.

For pooled WGS in **Paper I and III**, TruSeq DNA kits (Illumina Inc., CA, USA.) were used for preparing DNA libraries according to the manufacturer’s instructions except for the following changes. An MBS 1200 pipetting station (Nordiag AB, Stockholm Sweden) was used for automating protocols and substitutive magnetic bead clean-up methods were used for purification and gel-cut (Borgstrom et al., 2011; Lundin et al., 2010). The clustering was performed on a cBot cluster generation system using a HiSeq paired-end read cluster generation kit (Illumina Inc.) The technical replication of each Bull Terrier pooled sample was performed WGS with a different DNA library preparation.

In **Paper I**, **II** and **III**, Pooled whole genome and exome sequencing were performed on an Illumina HiSeq 2000 (Illumina Inc.) as paired-end reads to 100 bp at SciLifeLab Core Facility (Stockholm, Sweden). The base conversion was done using OLB v1.9 (Illumina Inc.). In **Paper III**, the whole genome sequencing of two dyslexic male individuals was performed at Complete Genomics Inc. (CGI, CA, USA) using their protocols.

As described in the previous publication (Vesterlund et al., 2011), RiboMinus™ Eukaryote Kit (Thermo Fisher Scientific, MA, USA) was used for depleting rRNA from total RNA, and then RNA was fragmented by RNase III (Thermo Fisher Scientific) in **Paper IV**. Using 200 ng of fragmented RNA as input, RNA-Seq library was constructed according to the Small RNA Expression Kit protocol (Thermo Fisher Scientific), and sequenced using SOLiD Opti Fragment Library Sequencing kit Master Mix 50 chemistry (Thermo Fisher Scientific) giving 50 bp reads on a SOLiD System Sequencing platform version 3 plus (Thermo Fisher Scientific). We performed a total of three runs of RNA-seq, with two biological replicate runs and one technical replicate runs.

4.3 DATA ANALYSIS

4.3.1 Alignment of sequencing reads

Paired-end Illumina sequence reads from pooled human samples in **Paper I, III** and **III** were aligned to the reference genomes with Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) version 0.6.1 for each pool. We used the National Center for Biotechnology Information (NCBI) human reference genome build 37 (GRCh37/hg19) as the reference genome. The same mapping software was applied for mapping Bull Terrier whole genome sequencing reads with *Canis lupus familiaris* genome assembly CanFam3.1 as reference. SAMtools (Li et al., 2009) was used for PCR duplicate removal in each pool (the WES studies: version 0.1.18; the WGS study: version 0.1.19) (Figure 5). In **Paper I** and **III**, Genome Analysis Toolkit (GATK) (McKenna et al., 2010) version 2.7.2 was applied for local realignment and base recalibration, taking variants from dbSNP 137 as reference. For the whole genome sequencing on Complete Genomics platform in **Paper III**, the CGI production pipeline aligned original data to NCBI reference genome build 37 with Complete Genomics Analysis Tools (CGA™ Tools).

In **Paper IV**, Tophat (Trapnell et al., 2009) v2.0.4 that applies Bowtie (Langmead et al., 2009) v.0.12.8 was utilized for mapping color space reads generated by the SOLiD platform. Sequencing reads from three runs at four stages were individually aligned to the zebrafish reference genome danRer7/Zv9 assembly from Ensembl. All uniquely mapped reads with mapping quality (MQ) equal to or greater than 20 were applied for further analysis.

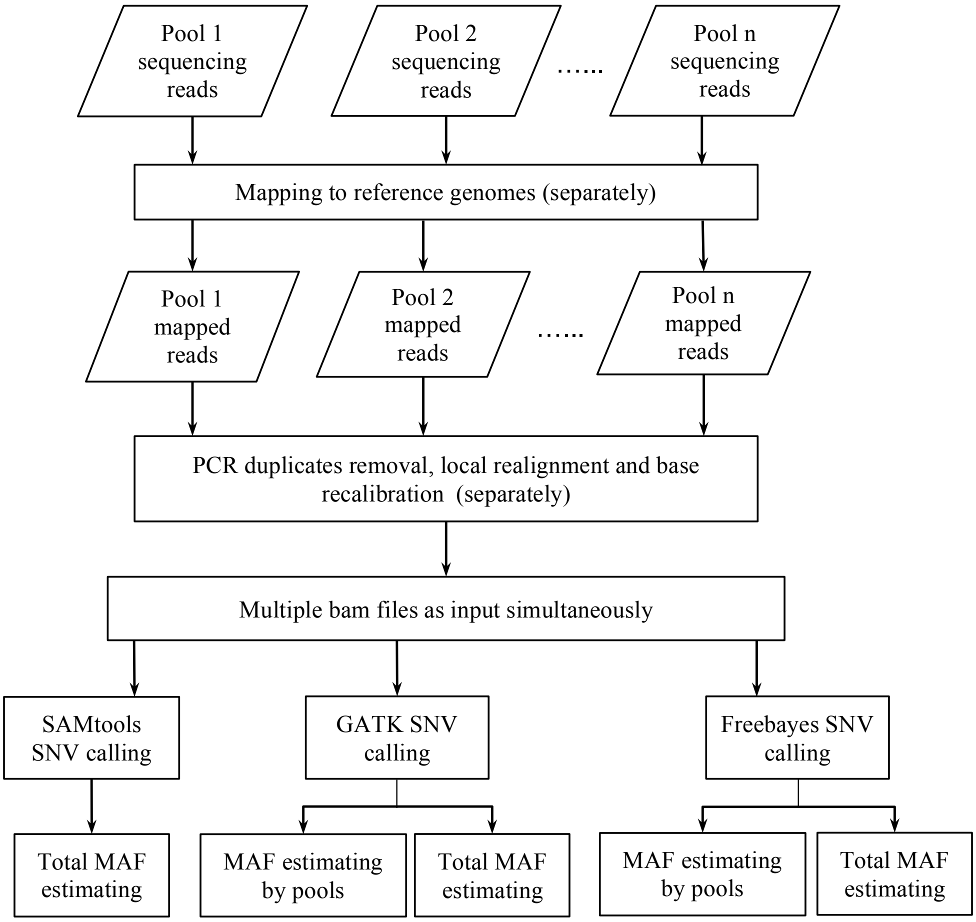
Pool 1 sequencing reads
Pool 2 sequencing reads
……
Pool n sequencing reads
Mapping to reference genomes (separately)
Pool 1 mapped reads
Pool 2 mapped reads
……
Pool n mapped reads
PCR duplicates removal, local realignment and base recalibration (separately)
Multiple bam files as input simultaneously
SAMtools SNV calling
GATK SNV calling
Freebayes SNV calling
Total MAF estimating
MAF estimating by pools
Total MAF estimating
MAF estimating by pools
Total MAF estimating

Figure 5. Workflow of evaluating pooling strategy with DNA sequencing data

4.3.2 Detection of genetic variations

In **Paper I**, we applied three programs: SAMtools mpileup function (version 0.1.19), GATK UnifiedGenotyper (UG) module (version 2.7.2 for the WES studies and version 3.2.2 for the WGS study) and Freebayes (Garrison and Marth, 2012) (version 0.9.21), to detect SNVs (Figure 5). Besides the default diploid setting in SAMtools and GATK UG, we also applied a ploidy setting of 20 when using GATK UG and Freebayes for detection. All uniquely aligned reads with high quality (MQ ≥ 20) from 10 pools in WES were simultaneously taken as input for variant detection.

To evaluate the detection of SNV from WGS in **Paper I**, the genotypes of 20 dogs from the Illumina array (in section 4.4.1) were taken as the true condition. The measurements used are defined as follows:

Sensitivity or true positive rate (TPR): $\frac{TP}{TP+FN}$

Specificity (SPC) or true negative rate: $\frac{TN}{FP+TN}$

Precision or positive predictive value (PPV): $\frac{TP}{TP+FP}$

Negative predictive value (NPV): $\frac{TN}{TN+FN}$

Accuracy (ACC): $\frac{TP+TF}{TP+FP+TN+FN}$

True positive (TP): the number of SNVs detected by both the WGS and the Illumina arrays; True negative (TN): the number of monomorphic loci that did not show SNVs in either the WGS or the Illumina array; False positive (FP): the number of SNVs detected by the WGS, but monomorphic in the Illumina array; False negative (FN): the number of SNVs detected by the Illumina array, but monomorphic in the WGS.

In **Paper II**, SAMtools mpileup (version 0.1.18) was employed for detecting variants in pooled exome sequencing samples with default setting. The uniquely mapped reads with high quality (MQ ≥ 20) from 10 pools were separately taken as input for calling variants in each pool. In **Paper III**, SNV and INDEL discovery was conducted in pooled samples using the default setting of GATK based on high quality data (MQ ≥ 20).

4.3.3 Evaluation of allele frequency estimates

In **Paper I**, when using GATK and Freebayes with ploidy setting, the alternative allele frequency of every variant in each pool was estimated based on allele counts. The minor allele counts from genotype information were extracted in each pooled sample, and then divided by total alleles in the pool (n=20). The total alternative allele frequency (AAF) was the ratio of alternative allele counts to total allele counts (n=200). In **Paper I** and **II**, another approach based on read depth was also utilized for estimating the allele frequency. SAMtools, GATK and Freebayes supplied coverage information at each variant site. The total AAF was the percentage of reads supporting the alternative allele across all 10 pools. (Figure 6)

For evaluating estimated allele frequency from WES in **Paper I**, root-mean-square deviation (RMSD) was employed to measure the difference between estimated MAF from exome sequencing and experimentally validated MAF by genotyping. The RMSD was calculated accordingly:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(MAF_i - estimated\_MAF_i)^2}$$

MAF: Experimentally validated minor allele frequency; estimated_MAF: minor allele frequency estimated from the exome sequencing data. In the Bull Terrier WGS study, taking genotypes from Illumina array as true condition, the allele frequency differences in pooled samples between the WGS and the Illumina array were calculated by directly comparing minor allele counts between the two platforms.

4.3.4 Annotation and filtering of variants

ANNOVAR (Wang et al., 2010) was applied for annotating the variants detected from pooled DNA sequencing data in **Paper I**, **II** and **III**. RefSeq, dbSNP and 1000 Genomes Project were the major data resources for variant annotation. Variants were usually categorized as common (AAF > 5%), low-frequency (AAF between 1% to 5%) and rare (AAF < 1%) according to the data from European population in 1000 Genomes project. Moreover, Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013) and UCSC Genome Browser (Meyer et al., 2012) were applied for variant visualization.

In **Paper II**, the steps for selecting candidate variants for validation:
(1) Filter out SNVs in only one IS pool;
(2) Filter out SNVs in more than one control pool of PE or obesity samples (Jiao et al., 2014);
(3) Select missense, nonsense, and splice site rare and low-frequency SNVs;
(4) Filtered out SNVs located in the "unreliable genes" in suspect gene list (Fajardo et al., 2012; Ju et al., 2011);
(5) Select SNVs on genes that were involved in skeleton, muscle or nervous system development or cell adhesion according to the Gene Ontology Consortium database (Ashburner et al., 2000);
(6) Additionally keep SNVs not shown in control pools by manual visualization.

In **Paper III**, the detected SNVs were filtered according to the following criteria:
(1) Select SNVs within the genomic area of the dyslexia susceptibility haplotype;
(2) Select SNVs with the same genotype in both individually sequenced samples;
(3) Select heterozygous SNVs shared in individually sequenced samples and pooled samples;
(4) Keep novel variants not annotated in dbSNP 137, 1000 Genomes Project (Phase 1 release) and SISu database (http://www.sisuproject.fi);
(5) Select the novel SNVs located in ROBO1 gene and 1 Mb upstream region.

INDEL filtration was similar to the strategy for SNVs, but an INDEL would be kept as long as it appeared in one of individuals and pooled samples.

The prediction of protein binding on intergenic variants was conducted using TRANSFAC Public database (Wingender et al., 2000), JASPAR database (Mathelier et al., 2013), UniPROBE database (Newburger and Bulyk, 2009), and P-Match software (Chekmenev et al., 2005).

4.3.5 Association analysis

The association analysis of candidate variants in **Paper II** were performed using PLINK (Purcell et al., 2007) with chi-square test. P value and odds ratio (OR) was calculated by PLINK. Several subgroups were classified for association analysis:
(1) Gender;
(2) Cobb angle: mild (10° -30°), moderate (31° - 44°) and severe (≥ 45°);
(3) Major curve: right thoracic and “all except right thoracic” (including “left thoracic”);
(4) Onset age: juvenile (4-9 years) and adolescent (10-20 years).

4.3.6 Identification of NTRs

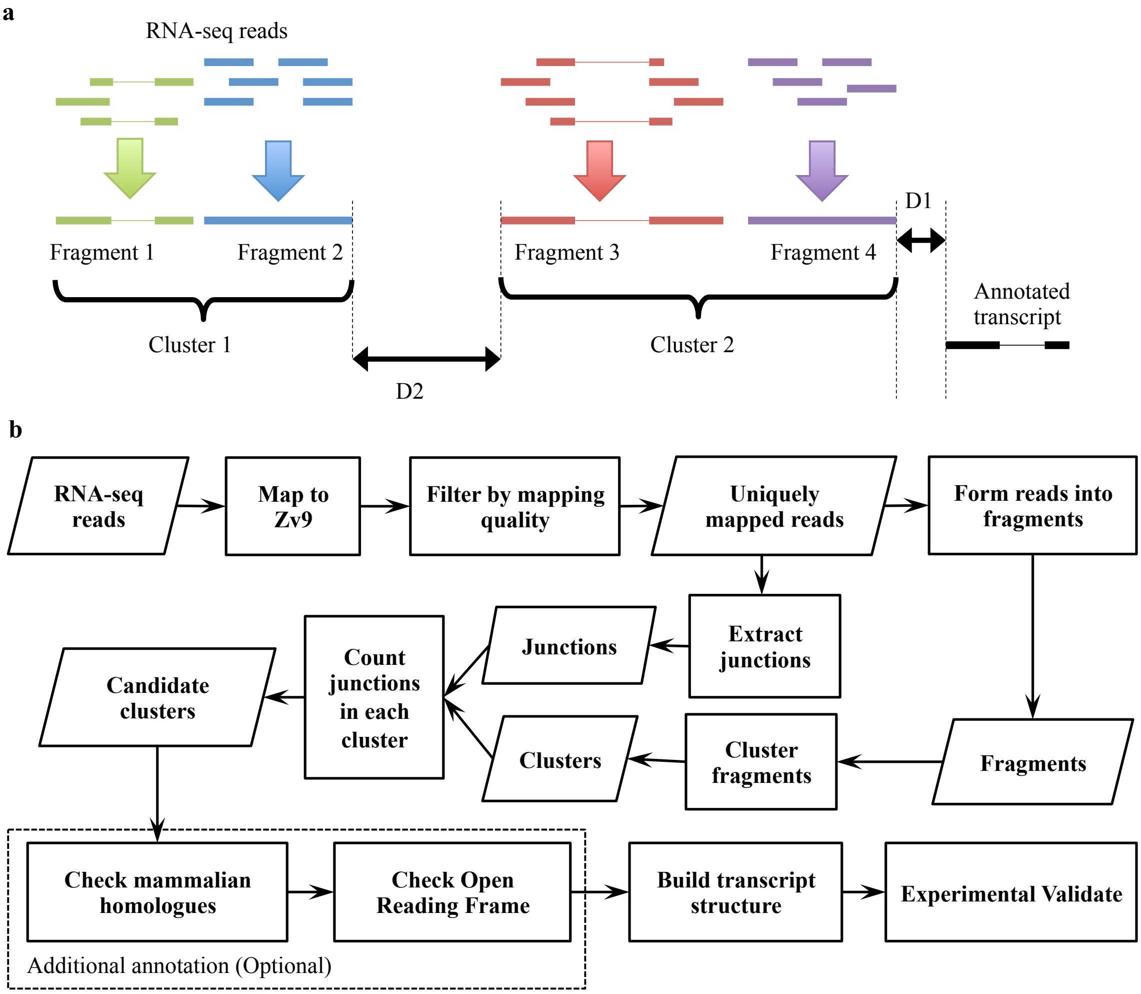
a
RNA-seq reads
Fragment 1
Fragment 2
Fragment 3
Fragment 4
D1
Annotated transcript
Cluster 1
Cluster 2
D2
b
RNA-seq reads
Map to Zv9
Filter by mapping quality
Uniquely mapped reads
Form reads into fragments
Extract junctions
Junctions
Count junctions in each cluster
Candidate clusters
Clusters
Cluster fragments
Fragments
Check mammalian homologues
Check Open Reading Frame
Build transcript structure
Experimental Validate
Additional annotation (Optional)

Figure 6. Workflow of putative NTR detection using RNA-seq data. a) construction of fragments and clusters. D1: the distance between a putative NTR and any annotated transcribed regions; D2: the distance between two putative NTRs. b) Systematic workflow of NTR identification.

In **Paper IV**, an in-house pipeline composed of BEDTools (Quinlan and Hall, 2010) modules and customized scripts was utilized for detecting NTRs. The pipeline performs two major tasks: constructing fragment from uniquely mapped reads and forming clusters (Figure 6b). A number of adjacent reads overlapped on the same strand were linked together to construct a fragment. A cluster was formed by a group of adjacent fragments within a certain distance (D2) on the same strand (Figure 6a). Parameters D1 and D2 were used for controlling the formation of clusters. D1 decided the distance between a formed cluster and any annotated transcript. D2 decided the maximum distance between two adjacent clusters. A cluster could be defined as a putative NTR if it meets two criteria: 1) It contained at least two splicing junction sites detected by TopHat; 2) It was not annotated in Ensembl genome annotation release 79. In addition, NCBI zebrafish (Danio rerio) annotation releases 103 and 104 were also employed for further filtering the putative NTRs. In order to confirm the putative protein-coding genes, models of genes provided by GENSCAN (Burge and Karlin, 1997) and ORF predicted using FGENESH software (http://www.softberry.com) were used. Several other *in silico* data were employed as additional supporting evidence of NTRs, e.g. zebrafish CAGE data (Nepal et al., 2013), conservation information supplied by tBLASTn and RefSeq gene of other species.

4.3.7 Gene expression

In **Paper IV**, the expression levels of the discovered NTRs at all four studied developmental stages were measured based on the RNA-seq data from the biological replicate samples. The Cuffdiff module from Cufflinks 2.0.0 (Trapnell et al., 2012) was utilized for estimating gene expression in terms of fragments per kilobase of transcript per million mapped reads (FPKM). The expression profiles were presented using Gplots package in R (http://CRAN.R-project.org/package=gplots) with the expression levels (FPKM) scaled in each NTR.

4.4 EXPERIMENTAL VALIDATION

4.4.1 Genotyping

4.4.1.1 Illumina SNP array

A group of tail-chasing Bull Terriers and controls were previously genotyped for GWAS. In **Paper I**, we utilized the genotyping data for validating the variants from WGS. The genotyping of 20 Bull Terriers included in the WGS was performed using Illumina Canine HD 173k SNP array (Illumina Inc.) by Illumina GenomeStudio (FIMM Technology Centre, Helsinki, Finland).

4.4.1.2 MassARRAY system

In **Paper I** and **II**, genotyping validation of human samples in the exome sequencing was conducted using iPLEX Gold chemistry and MassARRAY mass spectrometry system (Agena Bioscience, CA, USA). Multiplexed assays were designed using the MassARRAY Assay Design v4.0 Software (Sequenom, Agena Bioscience, CA, USA). Genotyping of the 100 IS samples was performed at the Mutation Analysis Facility (Karolinska University Hospital, Huddinge, Sweden), and genotyping of the 100 PE samples was performed at the Institute for Molecular Medicine Finland core facility (FIMM Technology Centre, University of Helsinki, Helsinki, Finland).

4.4.2 PCR

In **Paper II**, to capture an extra non-coding exon and potential TSSs of *LBX1* for detecting variants, primer pairs was designed using Primer3 (Koressaar and Remm, 2007; Untergasser et al., 2012). PCR amplification was performed on genomic DNA from the pooled IS samples with HotStarTaq Plus DNA Polymerase kit (Qiagen, Hilden, Germany). In Paper **III**, genomic DNA samples from dyslexic family members were used as the template for capturing the variants in *ROBO1* loci and *LHX2* binding site, with the same protocol of primer design and amplification described above.

4.4.3 RT-PCR

In **Paper IV**, four putative NTRs were selected for validation with RT-PCR. cDNA was synthesized using SuperScript III First-Strand Synthesis SuperMix according to the manufacturer's protocol (Thermo Fisher Scientific). To amplify regions of putative NTRs, HotStarTaq plus DNA polymerase kit (Qiagen) was used together with specific PCR primers designed by Primer 3 for covering the first and the last exon of predicted isoforms.

4.4.4 qRT-PCR

In **Paper IV**, 10 NTRs were randomly selected for validation of expression in four different developmental stages using qRT-PCR. Expression of bactin2 in each developmental stage was used as the control. The primers were designed to span the exon splicing junctions by using Primer3 mentioned above. According to the manufacturer's protocol, fast SYBR Green Master Mix (Thermo Fisher Scientific) was used for qRT-PCR and the experiments were run on the 7500 Real-Time PCR system (Applied Biosystems). Three biological replicates of each NTR in each developmental stage were used for quantitative measurement with b-actin as control. ΔCt was calculated as Ct of NTR minus Ct of b-actin. Average -ΔCt values were applied to demonstrate the expression patterns in 10 validated NTRs. ΔΔCt was calculated as ΔCt of each developmental stage minus ΔCt of 1-cell stage.

4.4.5 Sanger sequencing

In **Paper II** and **III**, all DNA amplicons were sequenced using cycle sequencing technology on an ABI 3730XL sequencing machine at Eurofin Genomics (Ebersberg, Germany). In **Paper IV**, cDNA fragments were subsequently cloned into pCRII-TOPO vector (Invitrogen, Thermo Fisher Scientific) and then sent for Sanger sequencing as described above.

4.4.6 Functional studies

In **Paper III**, an electrophoretic mobility shift assay (EMSA) was utilized for measuring protein and DNA binding affinity at *LHX2* binding site. We then employed Luciferase assay to evaluate the effect of variants on gene expression level compared with wild type and knocked down *LHX2* in lymphoblast cell lines to study the effect of *LHX2* on *ROBO1* gene expression. The protocols of those experiments were described in details in Paper III.

5 RESULTS

5.1 EVALUATION OF POOLING STRATEGY (PAPER I)

In this study, we evaluated the reliability of rare and low-frequency variant detection and accuracy of allele frequency estimates by using pooled DNA sequencing.

In the IS and the PE pools, Freebayes detected a higher proportion of singleton SNVs (SNVs in only one pool). SAMtools had the least amount of rare variants among three tools, but more common variants. When looking at the target enrichment regions, 99% of the SNVs detected by SAMtools were covered by GATK. Freebayes detected the largest number of SNVs in target regions, however the majority of the Freebayes-specific SNVs had a low possibility of being truly polymorphic, suggestion likely false positives. Thus it is critical to filter out the low quality SNVs when using Freebayes for variant detection. After conducting genotyping on individual samples, 42 SNVs and 44 SNVs selected from the IS and the PE studies respectively were successfully validated. Meanwhile we found four sites identified as polymorphic by variants calling software was actually monomorphic in experimental validation. In this case, SAMtools had the lowest false-positive detection rate, but it missed several sites that were actually polymorphic. GATK and Freebayes had similar performance in variant detection.

In each exome sequencing study, the total minor allele frequencies (MAFs) of SNVs estimated based on allele counts showed high similarities to the experimentally validated MAFs among all individuals (RMSD=0.031-0.032, Pearson correlation coefficient, *r*=0.88). Meanwhile, we also observed that the estimated MAFs were slightly exaggerated in the exome sequencing data. Taking the SNVs in the PE study as example, the total MAFs and those in each pool were overestimated for more than 50% of validated SNVs. The number of reads in each pool and the number of pools are found to affect the accuracy of MAF estimates. When excluding partial reads (20%-30%) in PE pool 7, three to four validated SNVs were not detected and the estimated MAFs of detected SNVs tended to be inflated. Randomly selecting PE pools, from one to nine pools, we calculated the difference between estimated and validated MAFs on a common SNV and a low-frequency SNV, and discovered that variant detection using more pools could lead to more precise estimates of MAFs.

In the Bull Terrier WGS study, GATK and Freebayes detected a similar number of SNVs and the SNVs discovered using those tools were highly overlapped (88-90%). Taking the genotyping of Illumina array as a reference, the performance measurements of GATK and Freebayes detection using WGS were over 96%, suggesting a high concordance between those two platforms at a genome-wide level. When using MAFs of autosomal markers in the Illumina array for MAF evaluation, the MAF estimated from the WGS and the arrays showed a good concordance rate (77%) in two pooled samples with r=0.94. When excluding autosomal monomorphic markers, 43% had identical MAFs and 41% had only one allele difference between two platforms. Especially for the low-frequency variants in the Illumina array, 56% of have identical MAFs and 26% have one allele difference between sequencing and individual genotyping data.

5.2 POOLED EXOME SEQUENCING AND CANDIDATE GENE FOR STUDYING IS (PAPER II)

Exome sequencing on pooled samples was conducted to identify susceptibility genes to IS, followed by an association validation in large cohort. The candidate gene study replicated the association of *LBX1* with IS in the Scandinavian cohort.

Approximately 307 to 412 million sequence reads were obtained in each pool, over 90% of which could be mapped to the human reference genome. Around 80-90% were covered by at least 30x reads. More than 1.7 million SNVs were detected from the pooled exome sequencing data. After variant filtering, 50 SNVs were selected for validation in 180 cases, including the previously exome sequenced subjects, and 245 controls, but seven of them failed in primer design or genotyping. The initial genotyping validated 42 SNVs and we selected 20 of them ($OR \geq 1.5$ or ≤ 0.67) for a follow-up genotyping in another 1,567 cases and 1,567 controls cohorts. Combining the cases and controls from two rounds of genotyping, we were unable to find genome-wide significant association of any of these to IS.

Among the four candidate variants reported from previous GWAS, only the common variant rs11190870, located downstream of *LBX1*, were validated to be strongly associated to IS in a Scandinavian case and control cohort ($p = 7\times10^{-18}$, $OR = 1.53$). The others demonstrated no or very weak association. The subgroup analyses showed rs11190870 had stronger association with right thoracic curve than others and with female than male. To further investigate the correlation between *LBX1* gene and IS, we performed Sanger sequencing in *LBX1* loci and looked into exome sequencing data. However, except a common innocent silent variant, there is no genetic variant in the coding region, 5'UTR, non-coding exon or the promoter regions of *LBX1* in pooled IS samples.

5.3 WHOLE GENOME SEQUENCING OF DYSLEXIA SUSCEPTIBILITY HAPLOTYPE (PAPER III)

The discovery of this study suggested a potential explanation about the segregation of dyslexia in a family.

WGS was utilized to investigate the genomic area surrounding *ROBO1* on individual and pooled samples. The individual WGS performed high coverage with average read depth at more than 50x on *ROBO1* region, while the coverage of pooled WGS was a bit shallow, with 24x average read depth on *ROBO1* and 1 Mb upstream region. We searched *ROBO1* and 1 Mb upstream region for unknown SNVs, INDELs and SVs. Combining the filtering and annotation, we discovered one intronic SNV, two intergenic SNVs and 38 INDELs that were not reported according to the public database. The novel intergenic INDELs appeared as typical microsatellite repeats, which are unlikely to have functional consequences. The intronic SNV was between the first non-coding exon and the first coding exon of the brain specific *ROBO1a* (NM_002941.3) splice variant, located in a potential enhancer region according to the FANTOM5 promoter atlas (Andersson et al., 2014; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014). Certain TFBSs were predicted around the unknown SNVs. However, EMSA analyses using RPE-1 cell nuclear extracts did not support transcription factor binding to the three unknown SNVs.

We further extended the search region to 5 Mb upstream of the *ROBO1* promoter region and discovered an unknown SNV situated in regulatory element: a TFBS for homeobox gene, including LIM homeobox 2 (*LHX2*). A reduced expression of *ROBO1* was observed after knocking down *LHX2* in the cell lines extracted from the family members and control individuals, which is consistent with the co-expression pattern of *LHX2* and *ROBO1* observed in FANTOM5 database. Furthermore, we investigated the binding of *LHX2* on the SNV site and evaluated the difference caused by alternative allele. The EMSA result suggested that transcription factor *LHX2* has a higher affinity for the alternative allele than for the reference allele. In addition, the luciferase assay showed that *LHX2* could bind to both the reference and the alternative allele, but it is hard to detect a significant difference between two alleles.

5.4 IDENTIFICATION OF NTRS USING RNA-SEQ (PAPER IV)

In this study, we identified *de novo* transcripts that have not been previously annotated in zebrafish early development, using relatively simple in-house bioinformatics pipelines.

By pooling 166 million uniquely mapping reads from all four stages and all three runs of RNA-seq, 487,937 fragments were formed with overlapped reads or reads containing splicing junction sites. To avoid the discovery of known transcripts or their potential extension, 296,628 fragments located in the reference transcriptome or less than 1 kb away were filtered out. The remaining 191,309 fragments formed 60,194 clusters with separating distances less or equal to 5 kb between fragments. After excluding cluster with less than two splicing junctions, there were 648 putative NTRs left for further investigation. During the course of the study, the zebrafish genome annotations were continuously updated. Thereafter, some of the putative NTRs acoording to Ensembl version 79 were subsequently annotated in NCBI zebrafish annotation, validating our approach for those genes. The putative NTRs were then checked against NCBI annotation. As a final result, 152 NTRs had not been previously annotated in the databases.

Four randomly selected NTRs were validated by experiment, showing high similarity with the predicted structures of NTRs. Certain isoforms were not detected in 50% epiboly, however they may exist during other studied developmental stages. In addition to confirming the predicted isoforms, novel isoforms with alternative splicing patterns of the NTR were also discovered.

The expression of the 152 NTRs was evaluated using RNA-seq data. Except for certian NTRs with high expression in 50% epiboly, more than 90% NTRs showed relatively low expression levels (FPKM < 1) in the four studied early developmental stages. Different expression patterns were discovered among them. Some NTRs demonstrated a clear upregulated pattern after 512-cell stage (MBT), which may be associated with processes important for organismal and anatomical structure development. Many NTRs demonstrated a dynamic change of expression during early development, with a peak in expression levels at the 512-cell stage (MBT). Other NTRs showed relatively high maternal expression were expressed throughout MBT to subsequently diminish in 50% epiboly stage.

Gene expression profiling by qRT-PCR was performed on 10 randomly selected NTRs. The expression measured by qRT-PCR replicated a similar expression profile for the 8 downregulated NTRs. The other two were detected as downregulated at the 50% epiboly stage by qRT-PCR, instead of large increases at 50% epiboly from RNA-seq data.

6 DISCUSSION

Pooling strategy demonstrated high reliability in detecting rare and low-frequency variants and high accuracy in estimating allele frequency. Despite the appealing feature of low cost for sequencing large number of samples, it is necessary to consider the limitations of this strategy when designing and analyzing pooled DNA sequencing (Schlötterer et al., 2014). One purpose for using pooling strategy is obtaining the MAFs among pooled samples. The unequal composition of individual DNA materials in a pool may result in inaccurate MAFs, especially at positions with shallow sequencing read depths. Moreover, sequencing errors are difficult to distinguish from rare or low-frequency alleles. Misalignments of divergent short reads are hard to identify in pooled samples and affect allele frequency estimates, especially for low-frequency variants. Special care should be taken when using current pooling protocol for detecting rare and low-frequency variants in association study. In addition, pooled DNA sequencing with short reads usually cannot supply haplotype information that is necessary for some gene burden tests depending on LD between rare variants.

In the IS study, more than 1 million SNVs were identified in coding regions from exome sequencing data. However, after filtering and genotyping validation, none of the selected functional variants were able to show significant association with IS. Aside from our selection, there were still large volumes of low-frequency missense variants not involved for association analysis. Beyond the regions covered by exome sequencing, the majority of the human genome was uninvestigated, so we could lose non-coding variants or long structure variations that may be linked to complex disorders (Alkan et al., 2011; Andersson et al., 2014). In addition, the limited sample size may affect the statistical power in rare or low-frequency variant association analysis. To reach sufficient statistic power for detecting rare variants in exome sequencing, 10,000 samples may be required (Kiezun et al., 2012). However, it would be a big challenge for sample collection, especially in a small population. Moreover, including more cases could introduce more noise from different sub-phenotypes.

By far, there are no other genetic variants reported to have equivalent significant association with IS as rs11190870. This association has been replicated in several East Asian, American and Scandinavian populations. Those discoveries indicate that rs11190870 may be a major risk factor for IS and the rest of variants may show very weak effects on IS. In such a case, what is a practical plan for further study?

One approach is to continue hunting the susceptibility rare or low-frequency variants, probably on a genome-wide scale, e.g. sequencing multiple affected individuals in a family. Subsequently screening variants based on their functions, population frequencies, functional consequence and co-segregation in the family could narrow down the candidates. A good example is the discovery of IS causative gene *POC5* by exome sequencing on three affected family members (Patten et al., 2015). However, the information described above may not be sufficient to distinguish the susceptibility genes in complex disorders. It would then need co-segregation evidence from additional family or association evidence from case-control further screening (Cirulli and Goldstein, 2010).

Another option is to perform functional study for characterizing the impact of genetic variants on the phenotype. Taking IS as example, how intergenic rs11190870 affects the phenotype is hard to determine even though it has strong association with IS. This common variant is located at a conservative domain, 7.5 kb downstream of *LBX1* that is a plausible cause for spinal deformity (Cheng et al., 2005; Gross et al., 2002; Jagla et al., 1995; Schäfer and Braun, 1999). The 3'-flanking region carrying rs11190870 was recently discovered to physically interact with the promoter region of *LBX1*. It worked like an enhancer that causes a higher transcriptional activity of *LBX1* in a human cell line (Guo et al., 2016). Guo et al. also performed overexpression of *lbx1* homologs on zebrafish embryos and the zebrafish demonstrated axial developmental defects. This case highlights the value of genome annotation for characterizing disease-associated variants in non-coding regions. When filtering candidate variants for validating association, those located in non-coding regions with limited understanding may be neglected. However, they are probably susceptible to disease, e.g. enhancers distally regulated through topological associated domain (Lupiáñez et al., 2015). Meanwhile, a more comprehensive annotation of model organism becomes vitally important for studying the function of genetic variants. The emphasis on genome annotation has moved to determining transposons, regulatory regions, pseudogenes and ncRNA genes (Yandell and Ence, 2012), which can be revealed and dissected with the assistance of various types of high-throughput sequencing.

7 CONCLUSION AND FUTURE PERSPECTIVE

Altogether, our studies utilized powerful high-throughput NGS to study the genetic basis of complex disorders and identify putative new transcripts for functional study in animal models. In **Paper I**, we observed that a large number of low-frequency variants were detected from the pooled DNA sequencing, and the estimated MAFs are highly concordance with the validated MAFs. Because of its reliability and accuracy, the pooling DNA sequencing could be a cost-efficient approach for initial screening in association study on large cohort. In **Paper II**, pooling strategy was implemented in exome sequencing for studying IS. However, we were unable to find evidence of a strong association in our candidate low-frequency variants to the phenotype. The rare variant discovered from pooled and individual whole genome sequencing in **Paper III** suggested a possible explanation for dyslexia in the family. Nevertheless, the potential roles of other genetic variations in the linkage regions cannot be excluded. In **Paper IV**, based on RNA-seq data, we identified 152 putative NTRs using an in-house pipeline. Randomly selected NTRs have been experimental validated and quantified during different developmental stages. The discovery of NTRs provided more information for characterizing the zebrafish genome and studying early development in model organisms.

High-throughput NGS of short reads have identified thousands of rare variants that could contribute to the genetic basis of complex disorders. The rare variants with large effects may partially explain the missing heritability. In addition, other underlying factors have also been suggested, such as structural variation, epistasis, epigenetics and shared environment (Maher, 2008; Manolio et al., 2009). One of the factors, large-scale structural variants, are not obvious to resolve by using current protocol of short DNA sequencing, especially in highly complex regions. The single molecule real-time sequencing would be able to sequence long reads of 10-200 kb spanning complex or repetitive regions. It has proven to be a more effective method for identifying disease-relevant structural variations (Ritz et al., 2014). Another risk factor accounting for missing inheritability may be epigenetic modifications that alter gene expression but do not change DNA sequencing (Trerotola et al., 2015). NGS techniques provide a high-resolution map of epigenetic structure and dynamics, even at the single cell level (Schwartzman and Tanay, 2015; Zentner and Henikoff, 2014).

For the past several years, high-throughput DNA sequencing has become more popular, and even become routine in genetic research. The advantage of rapid and high-throughput volumes at low cost makes the possibility of this technique as an approach for clinical use. To meet the increasing demands from clinics, NGS service providers have developed specific protocols such as Illumina's pan-cancer screening method and Qiagen's GeneReader (Goodwin et al., 2016). Targeted DNA sequencing panels are utilized for detecting somatic mutations in cancer genes and testing disease-causative genes on fetuses and individuals of reproductive age, subsequently guiding diagnosis and treatment therapy.

Large amounts of NGS data have been produced at a dramatic growth rate in both genetic research and clinics (Schatz and Langmead, 2013). To reduce space for data storage, several methods have been suggested: storing less, compression or cloud-based solutions. Since the sequencing prices have continued to drop for the past few years, reproducing whole genome or exome sequencing is probably a wise choice rather than spending a lot on storing huge datasets. The current mapping files in BAM format are still too large for storage, subsequently a new more compact format, CRAM, started to replace the BAM format. Cloud-based solution supply data upload at low cost, but analyses on cloud and data download are relatively expensive. Data formats differ among sequencing providers, so it requires standard format to facilitate NGS data sharing, as well as standardization of software for data analysis. Robust systems should be established to keep track of data security and control access to data, according to the informed consent obtained from donors and patients (Altman et al., 2016).

When NGS is employed in clinics to identify novel causal variants for rare diseases and assess the risk of individuals for common diseases, integrative methods and bioinformatic approaches for precise discovery are required for translating and interpreting (Altman et al., 2016; Lelieveld et al., 2016). The error profile for each sequencing platform should be evaluated for understanding their strengths and limitations, subsequently guiding the selection of platforms in terms of expected use. Nowadays, SMRT long-read sequencing has higher error rates than short-read sequencing (Goodwin et al., 2016). A combination of different platforms could be an option to complement each other. Furthermore, in order to decipher complicated NGS data, reference datasets would help recognize important genetic markers for potential clinical use. Currently, researchers are devoted to construct a diversity of disease-specific and population-based databases, such as ClinGen (Rehm et al., 2015), ClinVar (Landrum et al., 2016), ESP (http://evs.gs.washington.edu/EVS) and ExAC (http://exac.broadinstitute.org), and etc. These resources can also provide references to develop methods for analyzing admixed genomes and identifying medically important loci.

8 ACKNOWLEDGEMENTS

The research projects included in the thesis were performed at the Department of Biosciences and Nutrition (BioNut) and the Science for Life Laboratory (Scilifelab), Karolinska Institutet, and collaborated with several other research groups in Sweden and Finland. It was my pleasure to work together with a group of talented people for the past five years. I would like to express my sincere appreciation here to all of my colleagues and collaborators who supported me in this journey, especially to:

Hong Jiao, my principal supervisor who led me into the wonderful field of genetics. I am so blessed to have you as my supervisor. Thank you for your guidance, supervision, commendations, criticism, encouragement, advice, assistance and protection. Sometimes I felt you were like a mother caring for my academic life. I have learned a lot from you, not only in knowledge, but also in attitude towards scientific research. One lesson that cannot be ignored is critical thinking. You often addressed questions to me during our discussions, which inspired me to make rational and reasoned judgments more independently. Regardless of whether I stay in academia or not, I will benefit from this invaluable lesson in life.

Juha Kere, my co-supervisor, who gave me this opportunity to conduct my doctoral study here. I am honored to have been recruited into JKE group and been engaged in NGS studies. I would like to thank you for sharing your deep understanding on genetic research, brilliant ideas on new approaches and optimistic attitude towards research and life in every circumstance, even in crisis. Your extensive knowledge and interest, from stenography to haiku, have left a deep impression on me. **Thomas Svensson**, my co-supervisor, former superior. I would like to express my gratitude for recommending me to Juha and your continued support. Thanks for participating in my registration seminar as the only supervisor who was available at that moment. I will always remember your advice after that seminar: take a deep breathing and slow down. I still keep trying to do so.

Liselotte Vesterlund, my co-author, for your hard work on the zebrafish NTR project, including pipetting and cloning numerous of samples. I admire your sharp eyes on how to pick good primers and also on different scientific topics. I am really impressed by your enthusiasm for research. Your good ideas and wise decisions helped us form our zebreafish manuscript. Now I hope our second zebrafish manuscript will be delivered in the future. **Tiina Skoog**, my co-author, for your arrangement of NGS experiments, communicating with NGS service providers and managing all the pooled samples in our studies. When you answered the office phone, I subconsciously learned how to be kind and patient. It was very nice to share the same office with both of you and have interesting chats, even though Hong and I sometimes left you two alone in Huddinge.

Elisabet Einarsdottir, my co-author, for your expertise in genetic research. Thank you for your ingenious ideas in scoliosis projects. I appreciated the limited time at your lab bench preparing for PCR and DNA extraction, but probably do not want to go back to the lab again. As a close colleague on the JKE genetics team, you gave a lot of support and trust in data analysis. **Anna Grauers** and **Paul Gerdhem**, my co-authors, for your clinical experience in idiopathic scoliosis and considerate thoughts on designing the study. From recruiting patients to writing the manuscript, the study would not have been published without your effort. **Tea Kaartokallio**, my co-author, for your dedicated work on preeclampsia study and ingenious advice on the pooling strategy manuscript. I am grateful for our conversations in Skype meetings and long email discussions. **Satu Massinen**, **Andrea Bieder** and **Isabel Tapia Paez**, for your vast knowledge on dyslexia and hard work studying the function of dyslexia candidate genes. **Hannele Laivuori** and **Hannes Lohi**, for supplying the materials in pooled DNA sequencing and sharing your deep understanding on you research fields. The **co-authors** of four constitute papers, as well as the **co-authors** of the PE paper that is not included in the thesis, for your contribution to the studies, including sample collection, commenting and revising our manuscripts. Without your efforts, the studies cannot be accomplished. In addition, **Eira Leinonen**, for sorting out the consistent IDs of dyslexia family members between different systems, and **Osmo Hakosalo,** for your assistance in Bull Terrier genotyping data analysis.

I would like to acknowledge the staff at the Mutation Analysis Facility for your work on genotyping validation in our association study: **Kristina**, **Päivi**, **Astrid**, **Malin**, **Ann-Charlotte**, **Gunnar**, **Cecilia** and **Mosekunola**, In addition, there are many current and former colleagues in the JKE group who encouraged me and assisted in my PhD: **Gayathri**, **Nancy**, **Shintaro**, **Cilla**, **Hans**, **Linda**, **Suvi**, **Kaarel**, **Aparna**, **Mariann**, **Nathalie**, **Lovisa**, **Helena**, **Maria**, **Kristiina**, **Amitha**, **Pauliina**, **Virpi**, **Ettore** and **Morana**. I want to express my appreciation to **Ingegerd**, for your assistance in DNA preparation for NGS experiments, **Myriam**, for helping me collect the ethical permits of the dyslexia studies, **Elo**, for your companionship, encouragement, support and advice. I will remember our happy moments together at your home and in the forest for picking berries and mushrooms. **Eeva-Mari**, for your good suggestion on the time frame of thesis writing, and **Debora**, for praying together with me when I felt depressed and frustrated. May God grant you wisdom and establish the work of your hands.

I also want to thank the administrators at BioNut, especially **Erik Lundgren,** for saving my data on external hard drives, **Vivian Saucedo Hildebrand,** for dealing with my resident permit, and **Monica Ahlberg,** for your assistance in my PhD education, especially for your effort to assigning me to Månen lecture hall.

Thanks to my friends who are working or have worked at Karolinska Institutet: **David Brodin**, **Wen Cai**, **Hui Gao**, **Hong Jin**, **Xuan Li, Milica Putnik**, **Agata Smialowska, Lois Tang**, **Yongtao Xue**, **Ying Zhao**, **Jian Zhu** and **Ting Zhuang**, and to the kind folks at Scilifelab, **Jun Wang**, **Yue Hu**, **Emanuela Henao Diaz, Simon Sundling** and **Yilin Liu**, for helping me in different ways and giving me valuable advices and suggestions. It was my pleasure to organize the Swedish Bioinformatics Workshop 2015 together with **Johannes Alneberg**, **Wenjing Kang**, **Marcel Sauerbier**, **Benjamín Sigurgeirsson**, **Matthew The** and **Yan Zhou**.

My deep gratitude for working on thesis book is given to **Vivian Tse,** for proofreading my first manuscript and the thesis for language, and to **Zidong Lin**, for your excellent skill in design and your creativity in illustrating the front cover.

I would like to say thank you to members of the Nordic Chinese Christian Church in Stockholm. Together with Pastors **Anthony Shum** and **Billy Lo**, they have provided me with a warm family in Sweden and my life is more enriched beyond science. Furthermore, particular appreciations should be given to four sisters who have shared my apartment and their lives with me over the past four years: **Lini Yin Olofsson**, for your creativity in daily life, **Zhiqin Lao**, for your spiritual support and prayers, **Ruiqing Ni**, for our discussions about doctoral studies and future careers, and **Xuan Zhao**, for taking care of my physical and spiritual needs, as well as your delicious food that stimulates my appetite and inspiration. Thank you for reading my thesis and giving me practical suggestions as an outsider. Moreover, I want to thank other brothers and sisters who are current or previous members of the overseas students fellowship at our church: **Ziquan Cao**, **Hanfeng Chen**, **Tianqi Chen**, **Enoch Cheung**, **Yi Gong**, **Naiqiang He**, **Shu Huang**, **Jimmy Hui**, **Kevin Lau**, **Yat Long Tsoi**, **Yonghui Xu**, **Zhihao Yang**, **Hongyu Zhang,** **Maggie Chan**, **Fengcai Chang,** **Jingtian Chen**, **Qi Chen**, **Zoie Cheung**, **Mengyin Hu**, **Yifan Hu**, **Meiyu Huang**, **Mandy Lam**, **Sheng Li**, **Christine Lo**, **Siu Tze Mak**, **Shengnan Nie**, **Lily Xichen Pang**, **Alison Siu**, **Bonnie Siu**, **Wenjie Shen**, **Yan Wang**, **Liwen Wu**, **Nanxi Xie** and **Manshu Zhao**, for your communion every Sunday, sharing our happiness and sorrows. Most of you are and will be spread all over the world, but I will always remember and cherish the moments with you. **Yafeng Zhu** and **Peiyao Zhang**, in addition to those mentioned above, thanks for organizing my dissertation party.

Nine years ago I chose Sweden for my master studies, partly because my father’s cousin, **Linda Hui-Min Li Norrlinder**, lives here. Consequently my life has been changed. Thank you and **Bengt Norrlinder** for treating me like a family member.

Last but not the least, my beloved father **Shuping Wang** and mother **Huafang Meng**, I would like to thank you for your unconditional support and love, even though you probably do not understand what I have done in for my doctoral studies.

I would also like to acknowledge the financial support mainly from the **Swedish Research Council**, as well as **KI resebidrag och Axel Hirsch resebidrag för kirurger** for giving me opportunity to participate in an international conference.

9 REFERENCES

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G.P., Lim, A.Y.M., Hajan, H.S., Collas, P., Bourque, G., et al. (2011). Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. *21*, 1328–1338.

Aanes, H., Østrup, O., Andersen, I.S., Moen, L.F., Mathavan, S., Collas, P., and Alestrom, P. (2013). Differential transcript isoform usage pre- and post-zygotic genome activation in zebrafish. BMC Genomics *14*, 331.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Altaf, F., Gibson, A., Dannawi, Z., and Noordeen, H. (2013). Adolescent idiopathic scoliosis. BMJ *346*, f2508.

Altman, R.B., Prabhu, S., Sidow, A., Zook, J.M., Goldfeder, R., Litwack, D., Ashley, E., Asimenos, G., Bustamante, C.D., Donigan, K., et al. (2016). A research roadmap for next-generation sequencing informatics. Sci. Transl. Med. *8*, 335ps10–ps335ps10.

Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find. Am. J. Hum. Genet. *69*, 936–950.

Andersen, M.O., Christensen, S.B., and Thomsen, K. (2006). Outcome at 10 years after treatment for adolescent idiopathic scoliosis. Spine *31*, 350–354.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M.T., Perloski, M., Liberg, O., Arnemo, J.M., Hedhammar, Å., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature *495*, 360–364.

Bakwin, H. (1973). Reading Disability in Twins. Dev. Med. Child Neurol. *15*, 184–187.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Bansal, V., Harismendy, O., Tewhey, R., Murray, S.S., Schork, N.J., Topol, E.J., and Frazer, K.A. (2010). Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res. *20*, 537–545.

Baschal, E.E., Wethey, C.I., Swindle, K., Baschal, R.M., Gowan, K., Tang, N.L.S., Alvarado, D.M., Haller, G.E., Dobbs, M.B., Taylor, M.R.G., et al. (2015). Exome Sequencing Identifies a Rare HSPG2 Variant Associated with Familial Idiopathic Scoliosis. G3 GenesGenomesGenetics *5*, 167–174.

Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A., and Schlötterer, C. (2013). A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in Drosophila melanogaster. PLOS Genet *9*, e1003534.

Bateman, R.J., Xiong, C., Benzinger, T.L.S., Fagan, A.M., Goate, A., Fox, N.C., Marcus, D.S., Cairns, N.J., Xie, X., Blazey, T.M., et al. (2012). Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. N. Engl. J. Med. *367*, 795–804.

Borgstrom, E., Lundin, S., and Lundeberg, J. (2011). Large Scale Library Generation for High Throughput Sequencing. PLoS ONE *6*.

Buchan, J.G., Alvarado, D.M., Haller, G.E., Cruchaga, C., Harms, M.B., Zhang, T., Willing, M.C., Grange, D.K., Braverman, A.C., Miller, N.H., et al. (2014). Rare variants in FBN1 and FBN2 are associated with severe adolescent idiopathic scoliosis. Hum. Mol. Genet. *23*, 5271–5282.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

Chandrasekar, G., Vesterlund, L., Hultenby, K., Tapia-Paez, I., and Kere, J. (2013). The Zebrafish Orthologue of the Dyslexia Candidate Gene DYX1C1 Is Essential for Cilia Growth and Function. PLoS ONE *8*.

Chekmenev, D.S., Haid, C., and Kel, A.E. (2005). P-Match: transcription factor binding site search by combining patterns and weight matrices. Nucleic Acids Res. *33*, W432–W437.

Cheng, L., Samad, O.A., Xu, Y., Mizuguchi, R., Luo, P., Shirasawa, S., Goulding, M., and Ma, Q. (2005). Lbx1 and Tlx3 are opposing switches in determining GABAergic versus glutamatergic transmitter phenotypes. Nat. Neurosci. *8*, 1510–1515.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. *106*, 19096–19101.

Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.

Collins, F.S., Guyer, M.S., and Chakravarti, A. (1997). Variations on a Theme: Cataloging Human DNA Sequence Variation. Science *278*, 1580–1581.

Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Consortium, T. 1000 G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Consortium, T. 1000 G.P. (2012a). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Consortium, T.E.P. (2012b). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

DeFries, J.C., Fulker, D.W., and LaBuda, M.C. (1987). Evidence for a genetic aetiology in reading disability of twins. Nature *329*, 537–539.

DePristo, M.A., Banks, E., Poplin, R.E., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Derrien, T., Vaysse, A., André, C., and Hitte, C. (2011). Annotation of the domestic dog genome sequence: finding the missing genes. Mamm. Genome *23*, 124–131.

Derti, A., Garrett-Engele, P., MacIsaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science *327*, 78–81.

Druley, T.E., Vallania, F.L.M., Wegner, D.J., Varley, K.E., Knowles, O.L., Bonds, J.A., Robison, S.W., Doniger, S.W., Hamvas, A., Cole, F.S., et al. (2009). Quantification of rare allelic variants from pooled genomic DNA. Nat. Methods *6*, 263–265.

Einarsdottir, E., Svensson, I., Darki, F., Peyrard-Janvid, M., Lindvall, J.M., Ameur, A., Jacobsson, C., Klingberg, T., Kere, J., and Matsson, H. (2015). Mutation in CEP63. Hum. Genet. *134*, 1239–1248.

Fagerheim, T., Raeymaekers, P., Tønnessen, F.E., Pedersen, M., Tranebjærg, L., and Lubs, H.A. (1999). A new gene (DYX3) for dyslexia is located on chromosome 2. J. Med. Genet. *36*, 664–669.

Fajardo, K.V.F., Adams, D., Mason, C.E., Sincan, M., Tifft, C., Toro, C., Boerkoel, C.F., Gahl, W., and Markello, T. (2012). Detecting false positive signals in exome sequencing. Hum. Mutat. *33*, 609–613.

Fan, Y.-H., Song, Y.-Q., Chan, D., Takahashi, Y., Ikegawa, S., Matsumoto, M., Kou, I., Cheah, K.S., Sham, P., Cheung, K.M., et al. (2012). SNP rs11190870 near LBX1 is associated with adolescent idiopathic scoliosis in southern Chinese. J. Hum. Genet. *57*, 244–246.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Lassmann, T., Itoh, M., Summers, K.M., Suzuki, H., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. *7*, 85–97.

Fisher, S.E., and DeFries, J.C. (2002). Developmental dyslexia: genetic dissection of a complex cognitive trait. Nat. Rev. Neurosci. *3*, 767–780.

Fisher, S.E., Francks, C., Marlow, A.J., MacPhie, I.L., Newbury, D.F., Cardon, L.R., Ishikawa-Brush, Y., Richardson, A.J., Talcott, J.B., Gayán, J., et al. (2002). Independent genome-wide scans identify a chromosome 18 quantitative-trait locus influencing dyslexia. Nat. Genet. *30*, 86–91.

Flanagan, J.M., Sheehan, V., Linder, H., Howard, T.A., Wang, Y.-D., Hoppe, C.C., Aygun, B., Adams, R.J., Neale, G.A., and Ware, R.E. (2013). Genetic mapping and exome sequencing identify 2 mutations associated with stroke protection in pediatric patients with sickle cell anemia. Blood *121*, 3237–3245.

Francks, C., MacPhie, I.L., and Monaco, A.P. (2002). The genetic basis of dyslexia. Lancet Neurol. *1*, 483–490.

Friend, A., DeFries, J.C., and Olson, R.K. (2008). Parental Education Moderates Genetic Influences on Reading Disability. Psychol. Sci. *19*, 1124–1130.

Gao, W., Peng, Y., Liang, G., Liang, A., Ye, W., Zhang, L., Sharma, S., Su, P., and Huang, D. (2013). Association between Common Variants near LBX1 and Adolescent Idiopathic Scoliosis Replicated in the Chinese Han Population. PLoS ONE *8*, e53234.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio.

Gatz M, Reynolds CA, Fratiglioni L, and et al (2006). Role of genes and environments for explaining alzheimer disease. Arch. Gen. Psychiatry *63*, 168–174.

Gerdhem, P., and Akesson, K. (2007). Rates of fracture in participants and non-participants in the Osteoporosis Prospective Risk Assessment study. J. Bone Joint Surg. Br. *89*, 1627–1631.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The International HapMap Project. Nature *426*, 789–796.

Gilger, J.W.P.D., Pennington, B.F.P.D., and DeFRIES, J.C.P.D. (1992). A Twin Study of the Etiology of Comorbidity: Attention-deficit Hyperactivity Disorder and Dyslexia. J. Am. Acad. Child *31*, 343–348.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Gorman, K.F., Julien, C., and Moreau, A. (2012). The genetic epidemiology of idiopathic scoliosis. Eur. Spine J. *21*, 1905–1919.

Grauers, A., Rahman, I., and Gerdhem, P. (2012). Heritability of scoliosis. Eur. Spine J. *21*, 1069–1074.

Grauers, A., Danielsson, A., Karlsson, M., Ohlin, A., and Gerdhem, P. (2013). Family history and its association to curve size and treatment in 1,463 patients with idiopathic scoliosis. Eur. Spine J. *22*, 2421–2426.

Grauers, A., Topalis, C., Möller, H., Normelli, H., Karlsson, M., Danielsson, A., and Gerdhem, P. (2014). Prevalence of Back Problems in 1069 Adults With Idiopathic Scoliosis and 158 Adults Without Scoliosis. Spine.

Grimes, D.T., Boswell, C.W., Morante, N.F.C., Henkelman, R.M., Burdine, R.D., and Ciruna, B. (2016). Zebrafish models of idiopathic scoliosis link cerebrospinal fluid flow defects to spine curvature. Science *352*, 1341–1344.

Gross, M.K., Dottori, M., and Goulding, M. (2002). Lbx1 Specifies Somatosensory Association Interneurons in the Dorsal Spinal Cord. Neuron *34*, 535–549.

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. *47*, 435–444.

Guo, L., Yamashita, H., Kou, I., Takimoto, A., Meguro-Horike, M., Horike, S., Sakuma, T., Miura, S., Adachi, T., Yamamoto, T., et al. (2016). Functional Investigation of a Non-coding Variant Associated with Adolescent Idiopathic Scoliosis in Zebrafish: Elevated Expression of the Ladybird Homeobox Gene Causes Body Axis Deformation. PLOS Genet *12*, e1005802.

Haller, G., Alvarado, D., Mccall, K., Yang, P., Cruchaga, C., Harms, M., Goate, A., Willing, M., Morcuende, J.A., Baschal, E., et al. (2015). A polygenic burden of rare variants across extracellular matrix genes among individuals with adolescent idiopathic scoliosis. Hum. Mol. Genet. ddv463.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kaariainen, H., and Kere, J. (2005). The Axon Guidance Receptor Gene ROBO1 Is a Candidate Gene for Developmental Dyslexia. PLoS Genet. *1*.

Hardy, J., and Singleton, A. (2009). Genomewide Association Studies and Human Disease. N. Engl. J. Med. *360*, 1759–1768.

Harlaar, N., Spinath, F.M., Dale, P.S., and Plomin, R. (2005). Genetic influences on early word recognition abilities and disabilities: a study of 7-year-old twins. J. Child Psychol. Psychiatry *46*, 373–384.

Hawke, J.L., Wadsworth, S.J., and DeFries, J.C. (2006). Genetic influences on reading difficulties in boys and girls: the Colorado twin study. Dyslexia *12*, 21–29.

Hayes, M., Gao, X., Yu, L.X., Paria, N., Henkelman, R.M., Wise, C.A., and Ciruna, B. (2014). ptk7 mutant zebrafish models of congenital and idiopathic scoliosis implicate dysregulated Wnt signalling in disease. Nat. Commun. *5*, 4777.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nat. Genet. *39*, 1522–1527.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Jagla, K., Dollé, P., Mattei, M.-G., Jagla, T., Schuhbaur, B., Dretzen, G., Bellard, F., and Bellard, M. (1995). Mouse Lbx1 and human LBX1 define a novel mammalian homeo□ gene family related to the Drosophila lady bird genes. Mech. Dev. *53*, 345–356.

Jiang, H., Qiu, X., Dai, J., Yan, H., Zhu, Z., Qian, B., and Qiu, Y. (2012). Association of rs11190870 near LBX1 with adolescent idiopathic scoliosis susceptibility in a Han Chinese population. Eur. Spine J. *22*, 282–286.

Jiao, H., Arner, P., Gerdhem, P., Strawbridge, R.J., Näslund, E., Thorell, A., Hamsten, A., Kere, J., and Dahlman, I. (2014). Exome sequencing followed by genotyping suggests SYPL2 as a susceptibility gene for morbid obesity. Eur. J. Hum. Genet.

Ju, Y.S., Kim, J.-I., Kim, S., Hong, D., Park, H., Shin, J.-Y., Lee, S., Lee, W.-C., Kim, S., Yu, S.-B., et al. (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat. Genet. *43*, 745–752.

Kaartokallio, T., Klemetti, M.M., Timonen, A., Uotila, J., Heinonen, S., Kajantie, E., Kere, J., Kivinen, K., Pouta, A., Lakkisto, P., et al. (2014). Microsatellite Polymorphism in the Heme Oxygenase-1 Promoter Is Associated With Nonsevere and Late-Onset Preeclampsia. Hypertension *64*, 172–177.

Katusic, S.K., Colligan, R.C., Barbaresi, W.J., Schaid, D.J., and Jacobsen, S.J. (2001). Incidence of Reading Disability in a Population-Based Birth Cohort, 1976–1982, Rochester, Minn. Mayo Clin. Proc. *76*, 1081–1092.

Kere, J. (2011). Molecular genetics and molecular biology of dyslexia. Wiley Interdiscip. Rev. Cogn. Sci. *2*, 441–448.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.

Koressaar, T., and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. Bioinforma. Oxf. Engl. *23*, 1289–1291.

Kou, I., Takahashi, Y., Johnson, T.A., Takahashi, A., Guo, L., Dai, J., Qiu, X., Sharma, S., Takimoto, A., Ogura, Y., et al. (2013). Genetic variants in GPR126 are associated with adolescent idiopathic scoliosis. Nat. Genet. *45*, 676–679.

Laivuori, H., Lahermo, P., Ollikainen, V., Widen, E., Häivä-Mällinen, L., Sundström, H., Laitinen, T., Kaaja, R., Ylikorkala, O., and Kere, J. (2003). Susceptibility loci for preeclampsia on chromosomes 2p25 and 9p13 in Finnish families. Am. J. Hum. Genet. *72*, 168–177.

Lamminmäki, S., Massinen, S., Nopola-Hemmi, J., Kere, J., and Hari, R. (2012). Human ROBO1 regulates interaural interaction in auditory pathways. J. Neurosci. Off. J. Soc. Neurosci. *32*, 966–971.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lelieveld, S.H., Veltman, J.A., and Gilissen, C. (2016). Novel bioinformatic developments for exome sequencing. Hum. Genet. 1–12.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The Diploid Genome Sequence of an Individual Human. PLOS Biol *5*, e254.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, W., Li, Y., Zhang, L., Guo, H., Tian, D., Li, Y., Peng, Y., Zheng, Y., Dai, Y., Xia, K., et al. (2016). AKAP2 identified as a novel gene mutated in a Chinese family with adolescent idiopathic scoliosis. J. Med. Genet. jmedgenet – 2015–103684.

Londono, D., Kou, I., Johnson, T.A., Sharma, S., Ogura, Y., Tsunoda, T., Takahashi, A., Matsumoto, M., Herring, J.A., Lam, T.-P., et al. (2014). A meta-analysis identifies adolescent idiopathic scoliosis association with LBX1 locus in multiple ethnic groups. J. Med. Genet. *51*, 401–406.

Luk, K.D.K., Lee, C.F., Cheung, K.M.C., Cheng, J.C.Y., Ng, B.K.W., Lam, T.P., Mak, K.H., Yip, P.S.F., and Fong, D.Y.T. (2010). Clinical Effectiveness of School Screening for Adolescent Idiopathic Scoliosis: A Large Population-Based Retrospective Cohort Study. Spine *35*, 1607–1614.

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D., and Lundeberg, J. (2010). Increased Throughput by Parallelization of Library Preparation for Massive Sequencing. PLoS ONE *5*.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell *161*, 1012–1025.

Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., et al. (2010). Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. N. Engl. J. Med. *362*, 1181–1191.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nat. News *456*, 18–21.

Majander, K.K., Villa, P.M., Kivinen, K., Kere, J., and Laivuori, H. (2013). A follow-up linkage study of Finnish pre-eclampsia families identifies a new fetal susceptibility locus on chromosome 18. Eur. J. Hum. Genet. EJHG.

Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M.J., and Seddon, J.M. (2006). Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. Nat. Genet. *38*, 1055–1059.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods *7*, 111–118.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Mardis, E.R. (2008). Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet. *9*, 387–402.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.-Y., Chou, A., Ienasescu, H., et al. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res.

Matsson, H., Huss, M., Persson, H., Einarsdottir, E., Tiraboschi, E., Nopola-Hemmi, J., Schumacher, J., Neuhoff, N., Warnke, A., Lyytinen, H., et al. (2015). Polymorphisms in DCDC2 and S100B associate with developmental dyslexia. J. Hum. Genet. *60*, 399–401.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

McGuigan, F.E., Larzenius, E., Callreus, M., Gerdhem, P., Luthman, H., and Akesson, K. (2007). Variation in the BMP2 gene: bone mineral density and ultrasound in young adult and elderly women. Calcif. Tissue Int. *81*, 254–262.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2012). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

Miyake, A., Kou, I., Takahashi, Y., Johnson, T.A., Ogura, Y., Dai, J., Qiu, X., Takahashi, A., Jiang, H., Yan, H., et al. (2013). Identification of a Susceptibility Locus for Severe Adolescent Idiopathic Scoliosis on Chromosome 17q24.3. PLoS ONE *8*.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M.M., Sheng, Y., Abdelhamid, R.F., Anand, S., et al. (2013). Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res. *23*, 1938–1950.

Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. *37*, D77–D82.

Ng, P.C., and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. Genome Res. *11*, 863–874.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.

Ng, S.B., Nickerson, D.A., Bamshad, M.J., and Shendure, J. (2010). Massively parallel sequencing and rare disease. Hum. Mol. Genet. *19*, R119–R124.

Nopola-Hemmi, J., Myllyluoma, B., Haltia, T., Taipale, M., Ollikainen, V., Ahonen, T., Voutilainen, A., Kere, J., and Widen, E. (2001). A dominant gene for developmental dyslexia on chromosome 3. J. Med. Genet. *38*, 658–664.

Nopola-Hemmi, J., Myllyluoma, B., Voutilainen, A., Leinonen, S., Kere, J., and Ahonen, T. (2002). Familial dyslexia: neurocognitive and genetic correlation in a large Finnish family. Dev. Med. Child Neurol. *44*, 580–586.

O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., MacKenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. *43*, 585–589.

Out, A.A., van Minderhout, I.J.H.M., Goeman, J.J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P.E.M., Tops, C.M.J., et al. (2009). Deep sequencing to reveal new variants in pooled DNA samples. Hum. Mutat. *30*, 1703–1712.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Paracchini, S. (2011). Dissection of genetic associations with language-related traits in population-based cohorts. J. Neurodev. Disord. *3*, 365–373.

Patten, S.A., Margaritte-Jeannin, P., Bernard, J.-C., Alix, E., Labalme, A., Besson, A., Girard, S.L., Fendri, K., Fraisse, N., Biot, B., et al. (2015). Functional variants of POC5 identified in patients with idiopathic scoliosis. J. Clin. Invest. *125*, 1124–1128.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. *26*, 841–842.

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res. *30*, 3894–3900.

Ramus, F., and Szenkovits, G. (2008). What phonological deficit? Q. J. Exp. Psychol. *61*, 129–141.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen — The Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.

Risch, N., and Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. Science *273*, 1516–1517.

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., and Raphael, B.J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. Bioinformatics *30*, 3458–3466.

Rogala, E.J., Drummond, D.S., and Gurr, J. (1978). Scoliosis: incidence and natural history. A prospective epidemiological study. J. Bone *60*, 173–176.

Rutter M, Caspi A, Fergusson D, and et al (2004). Sex differences in developmental reading disability: New findings from 4 epidemiological studies. JAMA *291*, 2007–2012.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Schäfer, K., and Braun, T. (1999). Early specification of limb muscle precursor cells by the homeobox gene Lbx1h. Nat. Genet. *23*, 213–216.

Schatz, M.C., and Langmead, B. (2013). The DNA Data Deluge. IEEE Spectr. *50*, 26–33.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. *15*, 749–763.

Schwartzman, O., and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. Nat. Rev. Genet. *16*, 716–726.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA Pooling: a tool for large-scale association studies. Nat. Rev. Genet. *3*, 862–871.

Sharma, S., Gao, X., Londono, D., Devroy, S.E., Mauldin, K.N., Frankel, J.T., Brandon, J.M., Zhang, D., Li, Q.-Z., Dobbs, M.B., et al. (2011). Genome-wide association studies of adolescent idiopathic scoliosis suggest candidate susceptibility genes. Hum. Mol. Genet. *20*, 1456–1466.

Shaywitz, S.E. (1998). Dyslexia. N. Engl. J. Med. *338*, 307–312.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Stein, C.M., Schick, J.H., Gerry Taylor, H., Shriberg, L.D., Millard, C., Kundtz-Kluge, A., Russo, K., Minich, N., Hansen, A., Freebairn, L.A., et al. (2004). Pleiotropic Effects of a Chromosome 3 Locus on Speech-Sound Disorder and Reading. Am. J. Hum. Genet. *74*, 283–297.

Stevenson, J., Graham, P., Fredman, G., and Mcloughli, V. (1987). A Twin Study of Genetic Influences on Reading and Spelling Ability and Disability. J. Child Psychol. Psychiatry *28*, 229–247.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. Science *321*, 956–960.

Takahashi, Y., Kou, I., Takahashi, A., Johnson, T.A., Kono, K., Kawakami, N., Uno, K., Ito, M., Minami, S., Yanagida, H., et al. (2011). A genome-wide association study identifies common variants near LBX1 associated with adolescent idiopathic scoliosis. Nat. Genet. *43*, 1237–1240.

Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., and Li, J.B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Res. *23*, 201–216.

Tarkar, A., Loges, N.T., Slagle, C.E., Francis, R., Dougherty, G.W., Tamayo, J.V., Shook, B., Cantino, M., Schwartz, D., Jahnke, C., et al. (2013). DYX1C1 is required for axonemal dynein assembly and ciliary motility. Nat. Genet. *45*, 995–1003.

Terminology Committee of the Scoliosis Research Society (1976). A Glossary of Scoliosis Terms. Spine *1*, 57–58.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Tiira, K., Hakosalo, O., Kareinen, L., Thomas, A., Hielm-Björkman, A., Escriou, C., Arnold, P., and Lohi, H. (2012). Environmental Effects on Compulsive Tail Chasing in Dogs. PLOS ONE *7*, e41684.

Töhönen, V., Katayama, S., Vesterlund, L., Jouhilahti, E.-M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Bürglin, T.R., et al. (2015). Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat. Commun. *6*, 8207.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Trerotola, M., Relli, V., Simeone, P., and Alberti, S. (2015). Epigenetic inheritance and the missing heritability. Hum. Genomics *9*.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3--new capabilities and interfaces. Nucleic Acids Res. *40*, e115.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., and Kere, J. (2011). The zebrafish transcriptome during early development. BMC Dev. Biol. *11*, 30.

Visscher, P.M. (2008). Sizing up human height variation. Nat. Genet. *40*, 489–490.

Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat. Genet. *42*, 1109–1112.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Willner, S., and Udén, A. (1982). A prospective prevalence study of scoliosis in Southern Sweden. Acta Orthop. Scand. *53*, 233–237.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. *28*, 316–319.

Wynne-Davies, R. (1968). Familial (idiopathic) scoliosis. A family survey. J. Bone Joint Surg. Br. *50*, 24–30.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593–597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1131–1139.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342.

Yang, H., Zhou, Y., Gu, J., Xie, S., Xu, Y., Zhu, G., Wang, L., Huang, J., Ma, H., and Yao, J. (2013). Deep mRNA Sequencing Analysis to Capture the Transcriptome Landscape of Zebrafish Embryos and Larvae. PLOS ONE *8*, e64058.

Zentner, G.E., and Henikoff, S. (2014). High-resolution digital profiling of the epigenome. Nat. Rev. Genet. *15*, 814–827.