

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
13 February 2003 (13.02.2003)

PCT

(10) International Publication Number
WO 03/012139 A2

(51) International Patent Classification⁷: **C12Q 1/68**

(21) International Application Number: PCT/IB02/03347

(22) International Filing Date: 31 July 2002 (31.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/309,235 31 July 2001 (31.07.2001) US
60/316,830 31 August 2001 (31.08.2001) US

(71) Applicant and

(72) Inventor: **BOUGNERES, Pierre** [FR/FR]; Hospital Saint Vincent de Paul, Service d'Endocrinologie, 82, Rue Denfert Rochereau, F-75014 Paris (FR).

(74) Agents: **MARTIN, Jean-Jacques** et al.; Cabinet Regimbeau, 20, rue de Chazelles, F-75847 Paris Cedex 17 (FR).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU,

CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHODS FOR ASSESSING THE RISK OF OBESITY BASED ON ALLELIC VARIATIONS IN THE 5'-FLANKING REGION OF THE INSULIN GENE

(57) Abstract: The invention features methods for determining the risk of development of diabetes in a subject by examining the paternal insulin VNTR class. The invention further provides methods to facilitate rational therapy and maintenance of obese patients.



WO 03/012139 A2

METHODS FOR ASSESSING THE RISK OF OBESITY BASED ON ALLELIC VARIATIONS IN THE 5'-FLANKING REGION OF THE INSULIN GENE

FIELD OF THE INVENTION

5 The present invention relates to methods of diagnosis and treatment of obesity.

BACKGROUND OF THE INVENTION

For reasons that remain largely obscure, obesity is rapidly increasing in preschool children (1). Accumulation of excess fat in the first years of life is due to metabolic and hormonal events affecting the differentiation, proliferation and storage of lipids by adipocytes. Insulin is a potent regulator of fat accretion and neutral glyceride synthesis from glucose in early postnatal life (2). Sequence variations within the regulatory regions of the insulin gene (INS) have recently been shown to influence insulin secretion in children (3). Specifically, a polymorphic minisatellite located in the 5' region of INS influences the expression of both INS and the nearby insulin like growth factor 2 (IGF2) genes (4,5). During fetal life, genomic imprinting affects these two genes in humans, with restricted expression to the paternal allele. Paternal and maternal variable number of tandem repeats (VNTR)-INS-IGF2 haplotypes, therefore, do not have comparable roles during this period of life (6).

Caucasian INS VNTR alleles can be subdivided into two main length groups: class I (26-63 repeats) and class III (141-209 repeats). Class I alleles are associated with increased expression of INS in the fetal pancreas (7,8) and of IGF2 gene in the placenta (9). Several studies, in different control and diabetic populations, have shown departures from Mendelian parent-child transmission probabilities at this locus. In several Caucasian populations, Eaves et al found evidence for slight, but significant excess transmission of the class I allele from I/III heterozygous parents to healthy children (10). This transmission distortion was not specific to a particular parental gender, showing no evidence for parent-of-origin effects on excess transmission. However, two studies have shown parent-of-origin-dependent transmission distortion of VNTR alleles to children with Type 1 (T1D) or Type 2 diabetes (T2D). Bennett et al found an excess transmission of class I alleles from fathers to patients with autoimmune T1D (11). In contrast, Huxtable et al recently reported an excess transmission of class III alleles from fathers to T2D patients (12). This is a particularly interesting observation given that homozygous III/III individuals are known to have an increased risk of developing T2D (11).

Obesity and diabetes are among the most common human health problems in industrialized societies. In industrialized countries a third of the population is at least 20% overweight. In the United States, the percentage of obese people has increased from 25% at the end of the 70s, to 33% at the beginning of the 90's. Obesity is one of the most important risk factors for NIDDM. Definitions of obesity differ, but in general, a subject weighing at least 20% more than the recommended weight for his or her height and build is considered obese. The risk of developing

NIDDM is tripled in subjects 30% overweight, and three-quarters of people with NIDDM are overweight.

Obesity, which is the result of an imbalance between caloric intake and energy expenditure, is highly correlated with insulin resistance and diabetes in experimental animals and humans.

5 However, the molecular mechanisms that are involved in obesity-diabetes syndromes are not clear. During early development of obesity, increased insulin secretion balances insulin resistance and protects patients from hyperglycemia (Le Stunff, *et al.*, *Diabetes*.43, 696-702 (1994)). However, after several decades, β cell function deteriorates and non-insulin-dependent diabetes develops in about 20% of the obese population (Pedersen, P. *Diab. Metab. Rev.* 5, 505-509 (1989)) and (Brancati, 10 F.L., *et al.*, *Arch Intern Med.* 159, 957-963 (1999)). Given its high prevalence in modern societies, obesity has thus become the leading risk factor for NIDDM (Hill, J.O., *et al.*, *Science*. 280, 1371-1374 (1998)). However, the factors which predispose a fraction of patients to alterations of insulin secretion in response to fat accumulation remain unknown.

Obesity considerably increases the risk of developing cardiovascular diseases as well. Coronary 15 insufficiency, atheromatous disease, and cardiac insufficiency are at the forefront of the cardiovascular complications induced by obesity. It is estimated that if the entire population had an ideal weight, the risk of coronary insufficiency would decrease by 25%, and the risk of cardiac insufficiency and of cerebral vascular accidents by 35%. The incidence of coronary diseases is doubled in subjects under 50 years who are 30% overweight. The diabetic patient faces a 30% reduced lifespan. After age 45, 20 people with diabetes are about three times more likely than people without diabetes to have significant heart disease and up to five times more likely to have a stroke. These findings emphasize the inter-relations between risks factors for NIDDM and coronary heart disease and the potential value of an integrated approach to the prevention of these conditions based on the prevention of obesity (Perry, I.J. *et al.* *BMJ*. 310, 560-564 (1995)).

25 Despite advances in detecting mutations and genes associated with obesity, obesity continues to exert adverse effects on human health.

Literature

Bundred *et al.* (2001) *Brit Med.J.* **322**, 313-314; Taniguchi *et al.* (1986) *J. Lip. Res.* **27**, 925-929; Le Stunff *et al.* (2000) *Nat Genet.* **26**, 444-446; Kennedy *et al.* (1995) *Nat Genet.* **9**,293-298; 30 Paquette *et al.* (1998) *J. Biol. Chem.* **273**, 14158-64; Reik *et al.* (2001) *Nat. Rev.* **2**, 21-32; Vafiadis *et al.* (1996) *J. Autoimmun.* **9**,397-403; Bennett *et al.* (1996) *J. Autoimmun.* **9**, 415-421; Paquette *et al.* (1998) *J. Biol. Chem.* **273** :14158-14164; Eaves *et al.* (1999) *Nat. Genet.* **22**,324-5; Bennett and Todd (1996) *Annu. Rev. Genet.* **30**, 343-370 ; Huxtable *et al.* (2000) *Diabetes* **49**, 126-130.

35 SUMMARY OF THE INVENTION

The invention features methods for determining the risk of development of obesity by determining the insulin VNTR allele of the individual, particularly the paternal insulin VNTR allele.

In related aspects, the invention features methods to facilitate rational therapy and maintenance of individuals predisposed to become obese.

FEATURES OF THE INVENTION

The invention features a method of determining the risk of developing obesity in an individual. The method generally involves determining a paternal insulin VNTR allele in the individual. The presence of a paternal insulin VNTR class I allele indicates that the individual has an approximately two-fold increase in risk of developing obesity compared to an individual carrying a paternal insulin VNTR class III allele. Any method can be used to genotype the insulin VNTR in the individual, and thereby to determine the paternal insulin VNTR allele. In some embodiments, the determination is made by determining the identity of a polymorphic base of at least one marker in linkage disequilibrium with the insulin VNTR of the individual. In particular embodiments, the marker is -23 HphI.

The invention further features a method of treating obesity and related disorders in an individual. The method generally involves administering a weight loss or a weight control regimen in an individual identified by a method according to the invention as being at risk of developing obesity, thereby treating obesity in the individual. In some embodiments, a weight control regimen is selected from the group consisting of food restriction, increased calorie use, gastrointestinal surgery, medicinal approaches and reduced absorption of dietary lipids.

The invention further features a method of reducing the risk that an individual will develop an obesity-related disorder. The method generally involves administering a weight loss or a weight control regimen in an individual identified by a method according to the invention as being at risk of developing obesity, thereby reducing the risk that the individual will develop an obesity-related disorder.

DEFINITIONS

Before describing the invention in greater detail, the following definitions are set forth to illustrate and define the meaning and scope of the terms used to describe the invention herein.

The terms "insulin gene," when used herein, encompasses genomic, mRNA and cDNA sequences encoding the polypeptide hormone insulin, including the untranslated regulatory regions of the genomic DNA.

The term "isolated" requires that the material be removed from its original environment (e. g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition, and still be isolated in that the vector or composition is not part of its natural environment.

The term "isolated" further requires that the material be removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide present in a living animal is not isolated, but the same polynucleotide, separated from some or all of the coexisting materials in the natural system, is isolated. Specifically excluded from the definition of "isolated" are: naturally-occurring chromosomes (such as chromosome spreads), artificial chromosome libraries, genomic libraries, and cDNA libraries that exist either as an in vitro nucleic acid preparation or as a transfected/transformed host cell preparation, wherein the host cells are either an in vitro heterogeneous preparation or plated as a heterogeneous population of single colonies. Also specifically excluded are the above libraries wherein a specified polynucleotide of the present invention makes up less than 5% of the number of nucleic acid inserts in the vector molecules. Further specifically excluded are whole cell genomic DNA or whole cell RNA preparations (including said whole cell preparations which are mechanically sheared or enzymatically digested). Further specifically excluded are the above whole cell preparations as either an in vitro preparation or as a heterogeneous mixture separated by electrophoresis (including blot transfers of the same) wherein the polynucleotide of the invention has not further been separated from the heterologous polynucleotides in the electrophoresis medium (e.g., further separating by excising a single band from a heterogeneous band population in an agarose gel or nylon blot).

The term "purified" does not require absolute purity; rather, it is intended as a relative definition. Purification of starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated. As an example, purification from 0.1 % concentration to 10 % concentration is two orders of magnitude. The term "purified polynucleotide" is used herein to describe a polynucleotide or polynucleotide vector of the invention which has been separated from other compounds including, but not limited to other nucleic acids, carbohydrates, lipids and proteins (such as the enzymes used in the synthesis of the polynucleotide), or the separation of covalently closed polynucleotides from linear polynucleotides. A polynucleotide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polynucleotide sequence and conformation (linear versus covalently close). A substantially pure polynucleotide typically comprises about 50%, preferably 60 to 90% weight/weight of a nucleic acid sample, more usually about 95%, and preferably is over about 99% pure. Polynucleotide purity or homogeneity is indicated by a number of means well known in the art, such as agarose or polyacrylamide gel electrophoresis of a sample, followed by visualizing a single polynucleotide band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art.

The term "polypeptide" refers to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude post-expression modifications of polypeptides, for

example, polypeptides which include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

The term "recombinant polypeptide" is used herein to refer to polypeptides that have been artificially designed and which comprise at least two polypeptide sequences that are not found as contiguous polypeptide sequences in their initial natural environment, or to refer to polypeptides which have been expressed from a recombinant polynucleotide.

The term "purified polypeptide" is used herein to describe a polypeptide of the invention which has been separated from other compounds including, but not limited to nucleic acids, lipids, carbohydrates and other proteins. A polypeptide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polypeptide sequence. A substantially pure polypeptide typically comprises about 50%, preferably 60 to 90% weight/weight of a protein sample, more usually about 95%, and preferably is over about 99% pure. Polypeptide purity or homogeneity is indicated by a number of means well known in the art, such as polyacrylamide gel electrophoresis of a sample, followed by visualizing a single polypeptide band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art.

Throughout the present specification, the expression "nucleotide sequence" may be employed to designate indifferently a polynucleotide or a nucleic acid. More precisely, the expression "nucleotide sequence" encompasses the nucleic material itself and is thus not restricted to the sequence information (i.e. the succession of letters chosen among the four base letters) that biochemically characterizes a specific DNA or RNA molecule.

As used interchangeably herein, the terms "nucleic acids", "oligonucleotides", and "polynucleotides" include RNA, DNA, or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form. The term "nucleotide" as used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. Although the term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar, for examples of analogous linking groups, purine, pyrimidines, and sugars see for example PCT

publication No. WO 95/04064. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art.

5 A “promoter” refers to a DNA sequence recognized by the synthetic machinery of the cell required to initiate the specific transcription of a gene.

A sequence which is “operably linked” to a regulatory sequence such as a promoter means that said regulatory element is in the correct location and orientation in relation to the nucleic acid to control RNA polymerase initiation and expression of the nucleic acid of interest.

10 As used herein, the term “operably linked” refers to a linkage of polynucleotide elements in a functional relationship. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. More precisely, two DNA molecules (such as a polynucleotide containing a promoter region and a polynucleotide encoding a desired polypeptide or polynucleotide) are said to be “operably linked” if the nature of the linkage between the two polynucleotides does not (1) result in the introduction of a frame-shift mutation or (2) interfere with
15 the ability of the polynucleotide containing the promoter to direct the transcription of the coding polynucleotide.

The term “primer” denotes a specific oligonucleotide sequence which is complementary to a target nucleotide sequence and used to hybridize to the target nucleotide sequence. A primer serves as an initiation point for nucleotide polymerization catalyzed by either DNA polymerase, RNA
20 polymerase or reverse transcriptase.

The term “probe” denotes a defined nucleic acid segment (or nucleotide analog segment, e.g., polynucleotide as defined herein) which can be used to identify a specific polynucleotide sequence present in samples, said nucleic acid segment comprising a nucleotide sequence complementary of the specific polynucleotide sequence to be identified.

25 The terms “trait” and “phenotype” are used interchangeably herein and refer to any visible, detectable or otherwise measurable property of an organism such as symptoms of, or susceptibility to a disease for example. Typically the terms “trait” or “phenotype” are used herein to refer to symptoms of, or susceptibility to a disease, a beneficial response to or side effects related to a treatment. Preferably, said trait can be, but not limited to, obesity related disorders and/or diabetes
30 mellitus.

The term “allele” is used herein to refer to variants of a nucleotide sequence. A biallelic polymorphism has two forms. Diploid organisms may be homozygous or heterozygous for an allelic form.

35 The term “heterozygosity rate” is used herein to refer to the incidence of individuals in a population which are heterozygous at a particular allele. In a biallelic system, the heterozygosity rate is on average equal to $2P_a(1-P_a)$, where P_a is the frequency of the least common allele. In order to be

useful in genetic studies, a genetic marker should have an adequate level of heterozygosity to allow a reasonable probability that a randomly selected person will be heterozygous.

The term “genotype” as used herein refers the identity of the alleles present in an individual or a sample. In the context of the present invention, a genotype preferably refers to the description of the genetic marker alleles present in an individual or a sample. The term “genotyping” a sample or an individual for a genetic marker involves determining the specific allele or the specific nucleotide carried by an individual at a genetic marker.

The term “mutation” as used herein refers to a difference in DNA sequence between or among different genomes or individuals which has a frequency below 1%.

The term “haplotype” refers to a combination of alleles present in an individual or a sample. In the context of the present invention, a haplotype preferably refers to a combination of genetic marker alleles found in a given individual and which may be associated with a phenotype.

The term “polymorphism” as used herein refers to the occurrence of two or more alternative genomic sequences or alleles between or among different genomes or individuals. “Polymorphic” refers to the condition in which two or more variants of a specific genomic sequence can be found in a population. A “polymorphic site” is the locus at which the variation occurs. A single nucleotide polymorphism is the replacement of one nucleotide by another nucleotide at the polymorphic site. Deletion of a single nucleotide or insertion of a single nucleotide also gives rise to single nucleotide polymorphisms. In the context of the present invention, “single nucleotide polymorphism” preferably refers to a single nucleotide substitution. Typically, between different individuals, the polymorphic site may be occupied by two different nucleotides.

The term “biallelic polymorphism” and “genetic marker” are used interchangeably herein to refer to a single nucleotide polymorphism having two alleles at a fairly high frequency in the population. A “genetic marker allele” refers to the nucleotide variants present at a genetic marker site. Typically, the frequency of the less common allele of the genetic markers of the present invention has been validated to be greater than 1%, preferably the frequency is greater than 10%, more preferably the frequency is at least 20% (i.e. heterozygosity rate of at least 0.32), even more preferably the frequency is at least 30% (i.e. heterozygosity rate of at least 0.42). A genetic marker wherein the frequency of the less common allele is 30% or more is termed a “high quality genetic marker”.

The invention also concerns markers in linkage disequilibrium with the insulin HphI locus. The term “marker in linkage disequilibrium with the insulin HphI locus” is used herein to relate to the genetic markers described in Table A; preferably markers -4217 PstI, -2221 MspI, -23 HphI, +1428 FokI, +11000 AluI and +32000 ApaI; or more preferably marker -23 HphI. The term “marker in linkage disequilibrium with the insulin HphI locus” may include any other marker that is in linkage disequilibrium with the insulin HphI locus that is known in the art; as well as any marker determined to be in linkage disequilibrium with the insulin HphI locus by methods described herein.

The location of nucleotides in a polynucleotide with respect to the center of the polynucleotide are described herein in the following manner. When a polynucleotide has an odd number of nucleotides, the nucleotide at an equal distance from the 3' and 5' ends of the polynucleotide is considered to be "at the center" of the polynucleotide, and any nucleotide immediately adjacent to the nucleotide at the center, or the nucleotide at the center itself is considered to be "within 1 nucleotide of the center." With an odd number of nucleotides in a polynucleotide any of the five nucleotides positions in the middle of the polynucleotide would be considered to be within 2 nucleotides of the center, and so on. When a polynucleotide has an even number of nucleotides, there would be a bond and not a nucleotide at the center of the polynucleotide. Thus, either of the two central nucleotides would be considered to be "within 1 nucleotide of the center" and any of the four nucleotides in the middle of the polynucleotide would be considered to be "within 2 nucleotides of the center", and so on. For polymorphisms which involve the substitution, insertion or deletion of 1 or more nucleotides, the polymorphism, allele or genetic marker is "at the center" of a polynucleotide if the difference between the distance from the substituted, inserted, or deleted polynucleotides of the polymorphism and the 3' end of the polynucleotide, and the distance from the substituted, inserted, or deleted polynucleotides of the polymorphism and the 5' end of the polynucleotide is zero or one nucleotide. If this difference is 0 to 3, then the polymorphism is considered to be "within 1 nucleotide of the center." If the difference is 0 to 5, the polymorphism is considered to be "within 2 nucleotides of the center." If the difference is 0 to 7, the polymorphism is considered to be "within 3 nucleotides of the center," and so on.

The term "upstream" is used herein to refer to a location which is toward the 5' end of the polynucleotide from a specific reference point.

The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another by virtue of their sequence identities in a manner like that found in double-helical DNA with thymine or uracil residues linked to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, L., *Biochemistry*, 4th edition, 1995).

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym from "complementary polynucleotide", "complementary nucleic acid" and "complementary nucleotide sequence". These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind.

As used herein, the term "a condition related to obesity," refers to a condition (also referred to herein as a "disease" or a "disorder"), which is a direct or indirect result of, obesity. It is also a condition that is symptomatic of obesity. It is also a condition that occurs as a consequence of obesity. In particular, it is a condition that occurs at a higher frequency in obese individuals, as compared with non-obese individuals. Conditions associated with obesity include, but are not limited to, hypertension; atherosclerosis; Type II diabetes; osteoarthritis; breast cancer; uterine cancer; colon cancer; and coronary artery disease.

The term "obesity," as used herein, refers to a condition associated with excessive caloric intake relative to energy output such that excessive body fat accumulates. A standard measurement of obesity is body-mass index (BMI), which is defined as weight in kilograms divided by the square of the height in meters. A BMI of about 18.5-24.9 is considered the normal range for humans. A BMI of greater than 25.0 is considered overweight. The World Health Organization further categorizes "overweight" into grades: Grade 1, BMI = 25.0 to 29.9 (where the popular description is "overweight"); Grade 2, BMI= 30.0 to 39.9 (where the popular description is "obese"); and Grade 3, BMI = ≥ 40 (where the popular description is "morbidly obese". Thus, as used herein, "an obese individual" is one having a BMI of 30.0 or greater, and "a non-obese individual is one having a BMI of 29.9 or less. The term "obesity" includes early onset obesity and late onset obesity.

The term "early onset of obesity," as used herein, refers to obesity that first occurs in a child of between 12-15 years of age, between 10-12 years of age, between 8-10 years of age, between 6-8 years of age, between 4-6 years of age, between 2-4 years of age, or between birth and 2 years of age. Late onset obesity generally refers to obesity that occurs after about 15 years of age.

The term "hypertension," as used herein, refers to a condition identified by a systolic blood pressure of about 140 mm Hg or higher, a diastolic blood pressure of about 90 mm Hg or greater, or both.

The term "insulin-related disorder" refers to any disorder known in the art in which insulin production, secretion or function (i.e., insulin resistance) is altered in an individual. The term "insulin-related disorder" particularly refers to insulin-dependent diabetes mellitus (IDDM or Type I diabetes), or non-insulin dependent diabetes mellitus (NIDDM or Type II diabetes), gestational diabetes, autoimmune diabetes, hyperinsulinemia, hyperglycemia, hypoglycemia, β -cell failure, insulin resistance, dyslipemias, atheroma and insulinoma. The term "insulin-related disorder" further refers to obesity and obesity related disorders such as obesity-related NIDDM, obesity-related atherosclerosis, heart disease, obesity-related insulin resistance, obesity-related hypertension, microangiopathic lesions resulting from obesity-related NIDDM, ocular lesions caused by microangiopathy in obese individuals with obesity-related NIDDM, and renal lesions caused by microangiopathy in obese individuals with obesity-related NIDDM.

The terms "agent acting on an insulin-related disorder" refers to a drug or a compound modulating the activity of insulin production, insulin secretion, insulin function, decreasing the body

weight of obese individuals, or treating an insulin-related condition selected from the group consisting of IDDM, NIDDM, gestational diabetes, autoimmune diabetes, hyperinsulinemia, hyperglycemia, hypoglycemia, β -cell failure, insulin resistance, dyslipemias, atheroma, insulinoma, obesity and obesity related disorders as defined herein.

5 The terms “response to an agent acting on an insulin-related disorder” refer to drug efficacy, including but not limited to ability to metabolize a compound, to the ability to convert a pro-drug to an active drug, and to the pharmacokinetics (absorption, distribution, elimination) and the pharmacodynamics (receptor-related) of a drug in an individual.

 The terms “side effects to an agent acting on an insulin-related disorder” refer to adverse
10 effects of therapy resulting from extensions of the principal pharmacological action of the drug or to idiosyncratic adverse reactions resulting from an interaction of the drug with unique host factors.

 The term “NIDDM” as used herein refers to non-insulin-dependent diabetes mellitus or Type II diabetes (the two terms are used interchangeably throughout this document). NIDDM refers to a condition in which there is a relative disparity between endogenous insulin production and insulin
15 requirements, leading to an elevated blood glucose.

 The term “weight loss regimen” as used herein refers to any treatment known in the art aimed at reducing body mass. Weight loss regimens include food restriction, increased calorie use, gastrointestinal surgery, medicinal approaches and reduced absorption of dietary lipids.

 A “biological sample” encompasses a variety of sample types obtained from an individual
20 and can be used in a diagnostic or monitoring assay. The definition encompasses blood and other liquid samples of biological origin, solid tissue samples such as a biopsy specimen or tissue cultures or cells derived therefrom and the progeny thereof. The definition also includes samples that have been manipulated in any way after their procurement, such as by treatment with reagents, solubilization, or enrichment for certain components, such as polynucleotides. The term “biological
25 sample” encompasses a clinical sample, and also includes cells in culture, cell supernatants, cell lysates, serum, plasma, amniotic fluid, chorionic villus, biological fluid, and tissue samples.

 The term “patient” as used herein refers to a mammal, preferably primates, most preferably humans that are in need of treatment. The term “in need of such treatment” as used herein refers to a judgment made by a physician in the case of humans that a patient requires treatment. This judgment
30 is made based on a variety of factors that are in the realm of a physician’s expertise, but that include the knowledge that the patient is ill, or will be ill, as the result of a condition that is treatable by the compounds of the invention.

 Similarly, the term “individual” as used herein refers to a mammal, particularly a primate, preferably a human that perceives a need to reduce body mass (or that someone perceives the need to
35 reduce body mass for). The term “perceives a need” refers to modulations (increases) in body mass that are typically below the cut-off for clinical obesity, although could also include clinical obesity. “Modulations in body mass” is defined above.

Before the present invention is further described, it is to be understood that this invention is not limited to particular embodiments described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only by the appended claims.

Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either both of those included limits are also included in the invention.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can also be used in the practice or testing of the present invention, the preferred methods and materials are now described. All publications mentioned herein are incorporated herein by reference to disclose and describe the methods and/or materials in connection with which the publications are cited.

It must be noted that as used herein and in the appended claims, the singular forms “a”, “and”, and “the” include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to “a haplotype” includes a plurality of such haplotypes and reference to “the method” includes reference to one or more methods and equivalents thereof known to those skilled in the art, and so forth.

The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides methods for determining the risk of development of obesity by determining the insulin gene VNTR allele of the individual, particularly the paternal insulin gene VNTR allele. The invention further provides methods to facilitate rational therapy and maintenance of individuals with a paternal class I VNTR allele.

The invention results from the discovery that individuals who inherit an insulin (INS) VNTR class I allele from their father are nearly twice as likely to develop early onset obesity. This excess transmission was not observed for maternal class I alleles. The inventors determined the INS VNTR genotype of young obese patients, their lean sibling whenever possible, and both parents. The inventors found an unexpectedly large excess of paternal transmission of class I versus class III INS VNTR alleles to obese children. INS VNTR polymorphism is associated with variations in the expression of neighboring insulin and insulin like growth factor 2 (IGF2) genes. Fetal expression of these genes is restricted to the paternal chromosome as a consequence of genomic imprinting. Increased *in utero* expression of paternal insulin or IGF2 genes, due to the presence of a class I VNTR allele, predisposes one to postnatal fat deposition. No transmission distortion was seen from either parent to the lean siblings of the obese children. Due to the high frequency of the class I insulin allele, approximately 65-70% of Caucasian fetuses receive a class I VNTR allele from their father. This is an example of a widespread polymorphism associated with a significant risk of a common multifactorial disease.

In some embodiments, the invention features a method of determining the risk of developing obesity in an individual, comprising: a) determining the VNTR class of an insulin gene of the individual; and b) assigning a risk value, based on said genotype, of developing obesity. In another aspect, the invention features a method of determining the risk of developing obesity in an individual, comprising: a) determining the VNTR class of an insulin gene of the individual; b) determining the VNTR class of an insulin gene of a parent of the individual; and c) assigning a risk value, based on said VNTR class, of developing obesity. In another aspect, the invention features a method of determining the risk of developing obesity in an individual, comprising: a) determining the VNTR class of an insulin gene of the individual; b) determining the VNTR class of an insulin gene of the father of the individual; and c) assigning a risk value, based on said VNTR class, of developing obesity.

In a further embodiment, the invention features a method of treatment or prophylaxis of obesity for an individual comprising a method of prognosis of the invention and administering a weight loss or weight control regimen, wherein said weight loss regimen is selected from the group consisting of food restriction, increased calorie use, gastrointestinal surgery, medicinal approaches and reduced absorption of dietary lipids.

METHODS OF ASSESSING RISK OF DEVELOPING OBESITY

The invention provides methods of determining the risk in an individual of developing obesity. The methods generally involve determining the genotype of the insulin (INS) VNTR alleles of the individual. The presence in the individual of a paternal VNTR class I allele indicates that the individual has an approximately two-fold increased probability of developing obesity.

Individuals who are the subject of the genotyping include unborn fetuses, neonates, infants, and toddlers, e.g. individuals from pre-birth to about two years of age, from about two to about four

years of age, from about four to about six years of age, from about six to about eight years of age, from about eight to about ten years of age, from about ten to about 12 years of age, or from about 12 to about 15 years of age. A biological sample that contains the individual's genomic DNA is taken from the individual, and the DNA contained within the sample is used for genotyping. The source of DNA can be fetal cells (e.g., in a sample of amniotic fluid or chorionic villus); or any biological sample from a neonate, infant, or toddler that contains genomic DNA from the individual.

In general, in addition to genotyping the individual, at least the mother of the individual is genotyped. Where the genotype of the individual indicates that the individual is INS VNTR class I/INS VNTR class III, and the mother of the individual is homozygous for INS VNTR Class III, there is no need to genotype the biological father of the individual. In some cases, it may also be necessary to determine the INS VNTR genotype of the biological father of the individual. Where both parents have a VNTR class I allele, a second marker may be used to determine whether the individual has a paternal or a maternal VNTR class I allele. Thus, haplotype analysis can be used to determine whether the VNTR class I allele is paternal or maternal. Various methods, including, e.g. allele mapping by MVR-PCR, are described below and can be used to genotype an individual for the INS VNTR allele, and to determine whether a VNTR class I allele is paternal or maternal.

Methods for Genotyping an Individual for the INS VNTR allele

A variety of methods can be used to genotype a biological sample for insulin VNTR alleles, all of which may be performed *in vitro*. Such methods of genotyping comprise determining the identity of a nucleotide at an insulin-related genetic marker site by any method known in the art. An insulin-related genetic marker is any marker in linkage disequilibrium with the insulin HphI locus. This includes any marker known in the art which is a surrogate for the VNTR in the insulin gene. A list of markers in linkage disequilibrium with the insulin HphI locus is provided in Table A, below. For example, the -23 HphI (+) alleles are in complete linkage disequilibrium with class I alleles of neighboring VNTR. INS VNTR can be tested by using -23 HphI as a surrogate marker. The -23 HphI(+) single nucleotide polymorphism (SNP) genotype can be determined by analysis of polymerase chain reaction (PCR) products, e.g., using INS04 and INS05 primers, as described in Example 1.

These genotyping methods can be performed on nucleic acid samples derived from a single individual or pooled DNA samples. Typically, genotyping is performed on a DNA sample from an individual.

Source of DNA for Genotyping

Any source of nucleic acids, in purified or non-purified form, can be utilized as the starting nucleic acid, provided it contains or is suspected of containing the specific nucleic acid sequence desired. DNA or RNA may be extracted from cells, tissues, body fluids and the like as described above. While nucleic acids for use in the genotyping methods of the invention can be derived from

any primate source, the test subjects and individuals from which nucleic acid samples are taken are generally understood to be human.

Amplification of DNA Fragments Comprising Genetic markers

Many genotyping methods, although not all, require the previous amplification of the DNA region carrying the genetic marker of interest. Such methods specifically increase the concentration or total number of sequences that span the genetic marker or include that site and sequences located either distal or proximal to it. Diagnostic assays may also rely on amplification of DNA segments carrying a genetic marker of the present invention. Amplification of DNA may be achieved by any method known in the art. Amplification techniques are described above in the section entitled, Amplification of the Insulin Gene.

Some of these amplification methods are particularly suited for the detection of single nucleotide polymorphisms and allow the simultaneous amplification of a target sequence and the identification of the polymorphic nucleotide as it is further described below.

The genetic markers as described above allows the design of appropriate oligonucleotides, which can be used as primers to amplify DNA fragments comprising the genetic markers discussed herein. Amplification can be performed using the primers described herein or any set of primers allowing the amplification of a DNA fragment comprising a genetic marker associated with the INS gene.

In some embodiments genotyping is performed using primers for amplifying a DNA fragment containing one or more genetic markers associated with an INS gene. Exemplary amplification primers are listed in Table A and Table B. It will be appreciated that the primers listed are merely exemplary and that any other set of primers which produce amplification products containing one or more genetic markers of the present invention.

The spacing of the primers determines the length of the segment to be amplified. In the context of the present invention, amplified segments carrying genetic markers can range in size from at least about 25 bp to 35 kbp. Amplification fragments from 25-3000 bp are typical, fragments from 50-1000 bp are preferred and fragments from 100-600 bp are highly preferred. It will be appreciated that amplification primers for the genetic markers may be any sequence which allow the specific amplification of any DNA fragment carrying the markers.

Table A

Marker/ Position	Primers	Annealing Temp	PCR product	Alleles	Enzyme	Method of detection
TH microsatellite -9000	TH1 TH2	60°C	106/110/114 118/ 122 bp			6% acrylamide gel
-4217 PstI	TH9B TH10B	60°C	236 bp	T/C	PstI (1 U)	2% agarose gel in 0.5X TBE

Marker/ Position	Primers	Annealing Temp	PCR product	Alleles	Enzyme	Method of detection
-2733	INS68R INS68C	60°C		A/C		ARMS
-2221 MspI	INS56 INS57	63°C	186 bp	C/T	MspI (1 U)	2% agarose gel in 0.5X TBE
-365						Southern blot pINS310
-23 HphI	INS04 INS05	65°C	441 bp		HphI (2.5 U)	2% agarose gel The 9 bp band is not detectable
+805 DraIII					DraIII	
+1127 PstI					PstI	
+1140	INS71 INS71RC	60°C		A/C		ARMS
+1355	INS69 INS69RC	66°C		T/C		ARMS
+1404						Fnu4H1
+1428 FokI	ins13 DS02	65.5°C	433 bp		FokI (1 U)	1% agarose gel in 0.5X TBE
+2331	INS73A INS41	60°C		A/T		ARMS
+2336	INS55 INS41	64°C	116/121 bp			4% agarose gel (5 bp deletion)
+3201 HaeII	IIRI9 IIRI2B	62°C		G/A	HaeII	
+3580 MspI	INS46 INS47	60°C		G/A	MspI	
+3688	INS74C INS74R	64°C		C/T		ARMS
+3839 AlwN1	INS44 INS45	64°C		A/G	AlwN1	
+11000 AluI IGF2 exon 3	IGF2-26 IGF2-27	64°C	91 bp	C/T	AluI (1 U)	3% agarose-1000 gel in 0.5X TBE The 6 bp band is not detectable
+ 32000 ApaI	ApaIF ApaIR	55°C	236 bp		ApaI (1 U)	2% agarose-1000 gel in 0.5X TBE

(Lucassen, A.M.et al. Nature Genet. 4, 305-310 (1993))

Table B

Table Genotyping Primers					
Position	Primers (5'-3')	Annealing Temp	PCR product	Enzyme	Method of detection
-4217 PstI	TH9B: TGACGCCAAGGACAAGCTCA (SEQ ID NO:1) TH10B: CCAGCAGCCCCAGTCCTGCA (SEQ ID NO:2)	60°C	236 bp	PstI (1 U)	2% agarose gel in 0.5X TBE
-2221 MspI	INS56: CACCAGCTGGCCTTCAAGGT (SEQ ID NO:3) INS57: GCTGGGCACTAACAAGGTGT (SEQ ID NO:4)	63°C	186 bp	MspI (1 U)	2% agarose gel in 0.5X TBE
-23 HphI	INS04: TCCAGGACAGGCTGCATCAG (SEQ ID NO:5) INS05: AGCAATGGGCGTTGGCTCA (SEQ ID NO:6)	65°C	441 bp	HphI (2.5 U)	2% agarose gel in 0.5X TBE The 9 bp band is not detectable
+1428 FokI	ins13: TAAAGCCCTGAACCAGC (SEQ ID NO:7) DS02: CAGCCCAGCCTCCCTCCACA (SEQ ID NO:8)	65.5°C	433 bp	FokI (1 U)	1% agarose gel in 0.5X TBE
+11000 AluI	IGF2-26: CCCAGGGGCCGAAGAGTCA (SEQ ID NO:9) IGF2-27: GCTGAGCTGGCAGCGATTCA (SEQ ID NO:10)	64°C	91 bp	AluI (1 U)	3% agarose-1000 gel in 0.5X TBE The 6 bp band is not detectable
+ 32000 ApaI	ApaIF: CTTGGACTTTGAGTCAAATTGG (SEQ ID NO:11) ApaIR: CCTCCTTGGTCTTACTGGG (SEQ ID NO:12)	55°C	236 bp	ApaI (1 U)	2% agarose-1000 gel in 0.5X TBE

Methods of Genotyping DNA samples for Genetic Markers

Any method known in the art can be used to genotype DNA samples for a polymorphism associated with obesity by identifying a polymorphism in a marker in linkage disequilibrium with the HphI locus of the INS gene. Since the genetic marker allele to be detected has been identified and specified in the present invention, detection will prove simple for one of ordinary skill in the art by employing any of a number of techniques. Many genotyping methods require the previous amplification of the DNA region carrying the genetic marker of interest. While the amplification of target or signal is often preferred at present, ultrasensitive detection methods which do not require amplification or sequencing are also encompassed by the present genotyping methods. Methods well-known to those skilled in the art that can be used to detect genetic polymorphisms include methods such as, conventional dot blot analyzes, single strand conformational polymorphism analysis (SSCP) described by Orita et al. (1989) Proc. Natl. Acad. Sci. U.S.A.86: 2776-2770, denaturing gradient gel electrophoresis (DGGE), heteroduplex analysis, mismatch cleavage detection, and other conventional techniques as described in Sheffield, V.C. et al. (1991) Proc. Natl. Acad. Sci. U.S.A. 49:699-706, White, M.B. et al. (1992) Genomics. 12:301-306, Grompe, M. (1993) Nature Genetics. 5:111-117. Another method for determining the identity of the nucleotide present at a particular polymorphic site employs a specialized exonuclease-resistant nucleotide derivative as described in U.S. Pat. No. 4,656,127.

Exemplary methods involve directly determining the identity of the nucleotide present at a genetic marker site by sequencing assay, allele-specific amplification assay, or hybridization assay. The following is a description of some exemplary methods. One method is the microsequencing

technique. The term “sequencing” is used herein to refer to polymerase extension of duplex primer/template complexes and includes both traditional sequencing and microsequencing.

1) Sequencing Assays

The nucleotide present at a polymorphic site can be determined by sequencing methods. In a preferred embodiment, DNA samples are subjected to PCR amplification before sequencing as described above.

Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. Sequence analysis allows the identification of the base present at the genetic marker site.

2) Microsequencing Assays

In microsequencing methods, the nucleotide at a polymorphic site in a target DNA is detected by a single nucleotide primer extension reaction. This method involves appropriate microsequencing primers which, hybridize just upstream of the polymorphic base of interest in the target nucleic acid. A polymerase is used to specifically extend the 3' end of the primer with one single ddNTP (chain terminator) complementary to the nucleotide at the polymorphic site. Next the identity of the incorporated nucleotide is determined in any suitable way.

Typically, microsequencing reactions are carried out using fluorescent ddNTPs and the extended microsequencing primers are analyzed by electrophoresis on ABI 377 sequencing machines to determine the identity of the incorporated nucleotide as described in EP 412 883. Alternatively capillary electrophoresis can be used in order to process a higher number of assays simultaneously.

Different approaches can be used for the labeling and detection of ddNTPs. A homogeneous phase detection method based on fluorescence resonance energy transfer has been described by Chen and Kwok (1997) *Nucleic Acids Research*, 25:347-353 and Chen et al. (1997) *Proc. Natl. Acad. Sci. USA*, 94(20):10756-10761, the disclosures of which are incorporated herein by reference in their entireties. In this method, amplified genomic DNA fragments containing polymorphic sites are incubated with a 5'-fluorescein-labeled primer in the presence of allelic dye-labeled dideoxynucleoside triphosphates and a modified Taq polymerase. The dye-labeled primer is extended one base by the dye-terminator specific for the allele present on the template. At the end of the genotyping reaction, the fluorescence intensities of the two dyes in the reaction mixture are analyzed directly without separation or purification. All these steps can be performed in the same tube and the fluorescence changes can be monitored in real time. Alternatively, the extended primer may be analyzed by MALDI-TOF Mass Spectrometry. The base at the polymorphic site is identified by the mass added onto the microsequencing primer (see Haff L. A. and Smirnov I. P. (1997) *Genome Research*, 7:378-388), the disclosures of which are incorporated herein by reference in their entireties.

Microsequencing may be achieved by the established microsequencing method or by developments or derivatives thereof. Alternative methods include several solid-phase

microsequencing techniques. The basic microsequencing protocol is the same as described previously, except that the method is conducted as a heterogenous phase assay, in which the primer or the target molecule is immobilized or captured onto a solid support. To simplify the primer separation and the terminal nucleotide addition analysis, oligonucleotides are attached to solid supports or are modified in such ways that permit affinity separation as well as polymerase extension. The 5' ends and internal nucleotides of synthetic oligonucleotides can be modified in a number of different ways to permit different affinity separation approaches, e.g., biotinylation. If a single affinity group is used on the oligonucleotides, the oligonucleotides can be separated from the incorporated terminator reagent. This eliminates the need of physical or size separation. More than one oligonucleotide can be separated from the terminator reagent and analyzed simultaneously if more than one affinity group is used. This permits the analysis of several nucleic acid species or more nucleic acid sequence information per extension reaction. The affinity group need not be on the priming oligonucleotide but could alternatively be present on the template. For example, immobilization can be carried out via an interaction between biotinylated DNA and streptavidin-coated microtitration wells or avidin-coated polystyrene particles. In the same manner oligonucleotides or templates may be attached to a solid support in a high-density format. In such solid phase microsequencing reactions, incorporated ddNTPs can be radiolabeled (Syvänen, *Clinica Chimica Acta* 226:225-236, 1994) or linked to fluorescein (Livak and Hainer, *Human Mutation* 3:379-385, 1994). The detection of radiolabeled ddNTPs can be achieved through scintillation-based techniques. The detection of fluorescein-linked ddNTPs can be based on the binding of anti-fluorescein antibody conjugated with alkaline phosphatase, followed by incubation with a chromogenic substrate (such as p-nitrophenyl phosphate). Other possible reporter-detection pairs include: ddNTP linked to dinitrophenyl (DNP) and anti-DNP alkaline phosphatase conjugate (Harju et al., *Clin. Chem.* 39/11 2282-2287 (1993)) or biotinylated ddNTP and horseradish peroxidase-conjugated streptavidin with o-phenylenediamine as a substrate (WO 92/15712). As yet another alternative solid-phase microsequencing procedure, Nyren et al. (*Analytical Biochemistry* 208:171-175 (1993)), described a method relying on the detection of DNA polymerase activity by an enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA).

Pastinen et al. (*Genome Research* 7:606-614, 1997), describes a method for multiplex detection of single nucleotide polymorphism in which the solid phase minisequencing principle is applied to an oligonucleotide array format. High-density arrays of DNA probes attached to a solid support (DNA chips) are further below.

3) Allele-Specific Amplification Assay Methods

In one aspect the present invention provides polynucleotides and methods to determine the allele of one or more genetic markers of the present invention in a biological sample, by allele-specific amplification assays. Methods, primers and various parameters to amplify DNA fragments

comprising genetic markers of the present invention are further described above in "Amplification of DNA Fragments Comprising Genetic Markers".

Allele Specific Amplification Primers

Discrimination between the two alleles of a genetic marker can also be achieved by allele specific amplification, a selective strategy, whereby one of the alleles is amplified without amplification of the other allele. This is accomplished by placing the polymorphic base at the 3' end of one of the amplification primers. Because the extension forms from the 3' end of the primer, a mismatch at or near this position has an inhibitory effect on amplification. Therefore, under appropriate amplification conditions, these primers only direct amplification on their complementary allele. Determining the precise location of the mismatch and the corresponding assay conditions are well within the ordinary skill in the art.

Ligation/Amplification Based Methods

The "Oligonucleotide Ligation Assay" (OLA) uses two oligonucleotides which are designed to be capable of hybridizing to abutting sequences of a single strand of a target molecule. One of the oligonucleotides is biotinylated, and the other is detectably labeled. If the precise complementary sequence is found in a target molecule, the oligonucleotides will hybridize such that their termini abut, and create a ligation substrate that can be captured and detected. OLA is capable of detecting single nucleotide polymorphisms and may be advantageously combined with PCR as described by Nickerson D.A. et al. (1990) Proc. Natl. Acad. Sci. U.S.A. 87:8923-8927. In this method, PCR is used to achieve the exponential amplification of target DNA, which is then detected using OLA.

Other amplification methods which are particularly suited for the detection of single nucleotide polymorphism include LCR (ligase chain reaction), Gap LCR (GLCR) which are described above in "Amplification of the insulin gene". LCR uses two pairs of probes to exponentially amplify a specific target. The sequences of each pair of oligonucleotides, is selected to permit the pair to hybridize to abutting sequences of the same strand of the target. Such hybridization forms a substrate for a template-dependant ligase. In accordance with the present invention, LCR can be performed with oligonucleotides having the proximal and distal sequences of the same strand of a genetic marker site. In one embodiment, either oligonucleotide will be designed to include the genetic marker site. In such an embodiment, the reaction conditions are selected such that the oligonucleotides can be ligated together only if the target molecule either contains or lacks the specific nucleotide that is complementary to the genetic marker on the oligonucleotide. In an alternative embodiment, the oligonucleotides will not include the genetic marker, such that when they hybridize to the target molecule, a "gap" is created as described in WO 90/01069. This gap is then "filled" with complementary dNTPs (as mediated by DNA polymerase), or by an additional pair of oligonucleotides. Thus at the end of each cycle, each single strand has a complement capable of serving as a target during the next cycle and exponential allele-specific amplification of the desired sequence is obtained.

Ligase/Polymerase-mediated Genetic Bit AnalysisTM is another method for determining the identity of a nucleotide at a preselected site in a nucleic acid molecule (WO 95/21271). This method involves the incorporation of a nucleoside triphosphate that is complementary to the nucleotide present at the preselected site onto the terminus of a primer molecule, and their subsequent ligation to a second oligonucleotide. The reaction is monitored by detecting a specific label attached to the reaction's solid phase or by detection in solution.

4) Hybridization Assay Methods

A preferred method of determining the identity of the nucleotide present at a genetic marker site involves nucleic acid hybridization. The hybridization probes, which can be conveniently used in such reactions, preferably include the probes defined herein. Any hybridization assay may be used including Southern hybridization, Northern hybridization, dot blot hybridization and solid-phase hybridization (see Sambrook, J., Fritsch, E.F., and T. Maniatis. (1989) *Molecular Cloning: A Laboratory Manual*. 2ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York).

Hybridization refers to the formation of a duplex structure by two single stranded nucleic acids due to complementary base pairing. Hybridization can occur between exactly complementary nucleic acid strands or between nucleic acid strands that contain minor regions of mismatch. Specific probes can be designed that hybridize to one form of a genetic marker and not to the other and therefore are able to discriminate between different allelic forms. Allele-specific probes are often used in pairs, one member of a pair showing perfect match to a target sequence containing the original allele and the other showing a perfect match to the target sequence containing the alternative allele. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Stringent, sequence specific hybridization conditions, under which a probe will hybridize only to the exactly complementary target sequence are well known in the art (Sambrook et al., 1989). Stringent conditions are sequence dependent and will be different in different circumstances. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Although such hybridizations can be performed in solution, it is preferred to employ a solid-phase hybridization assay. The target DNA comprising a genetic marker of the present invention may be amplified prior to the hybridization reaction. The presence of a specific allele in the sample is determined by detecting the presence or the absence of stable hybrid duplexes formed between the probe and the target DNA. The detection of hybrid duplexes can be carried out by a number of methods. Various detection assay formats are well known which utilize detectable labels bound to either the target or the probe to enable detection of the hybrid duplexes. Typically, hybridization duplexes are separated from unhybridized nucleic acids and the labels bound to the duplexes are then detected. Those skilled in the art will recognize that wash steps may be employed to wash away

excess target DNA or probe as well as unbound conjugate. Further, standard heterogeneous assay formats are suitable for detecting the hybrids using the labels present on the primers and probes.

Two recently developed assays allow hybridization-based allele discrimination with no need for separations or washes (see Landegren U. et al., *Genome Research*, 8:769-776,1998). The TaqMan assay takes advantage of the 5' nuclease activity of Taq DNA polymerase to digest a DNA probe annealed specifically to the accumulating amplification product. TaqMan probes are labeled with a donor-acceptor dye pair that interacts via fluorescence energy transfer. Cleavage of the TaqMan probe by the advancing polymerase during amplification dissociates the donor dye from the quenching acceptor dye, greatly increasing the donor fluorescence. All reagents necessary to detect two allelic variants can be assembled at the beginning of the reaction and the results are monitored in real time (see Livak et al., *Nature Genetics*, 9:341-342, 1995). In an alternative homogeneous hybridization-based procedure, molecular beacons are used for allele discriminations. Molecular beacons are hairpin-shaped oligonucleotide probes that report the presence of specific nucleic acids in homogeneous solutions. When they bind to their targets they undergo a conformational reorganization that restores the fluorescence of an internally quenched fluorophore (Tyagi et al., *Nature Biotechnology*, 16:49-53, 1998).

The polynucleotides provided herein can be used in hybridization assays for the detection of genetic marker alleles in biological samples. These probes are characterized in that they preferably comprise between 8 and 50 nucleotides, and in that they are sufficiently complementary to a sequence comprising a genetic marker of the present invention to hybridize thereto and preferably sufficiently specific to be able to discriminate the targeted sequence for only one nucleotide variation. The GC content in the probes of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %. The length of these probes can range from 10, 15, 20, or 30 to at least 100 nucleotides, preferably from 10 to 50, more preferably from 18 to 35 nucleotides. A particularly preferred probe is 25 nucleotides in length. Preferably the genetic marker is within 4 nucleotides of the center of the polynucleotide probe. In particularly preferred probes the genetic marker is at the center of said polynucleotide. Shorter probes may lack specificity for a target nucleic acid sequence and generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. Longer probes are expensive to produce and can sometimes self-hybridize to form hairpin structures. Methods for the synthesis of oligonucleotide probes have been described above and can be applied to the probes of the present invention.

By assaying the hybridization to an allele specific probe, one can detect the presence or absence of a genetic marker allele in a given sample. High-Throughput parallel hybridizations in array format are specifically encompassed within "hybridization assays" and are described below.

5) Hybridization to Addressable Arrays of Oligonucleotides

Hybridization assays based on oligonucleotide arrays rely on the differences in hybridization stability of short oligonucleotides to perfectly matched and mismatched target sequence variants. Efficient access to polymorphism information is obtained through a basic structure comprising high-density arrays of oligonucleotide probes attached to a solid support (e.g., the chip) at selected positions. Each DNA chip can contain thousands to millions of individual synthetic DNA probes arranged in a grid-like pattern and miniaturized to the size of a dime.

The chip technology has already been applied with success in numerous cases. For example, the screening of mutations has been undertaken in the BRCA1 gene, in *S. cerevisiae* mutant strains, and in the protease gene of HIV-1 virus (Hacia et al., *Nature Genetics*, 14(4):441-447, 1996; Shoemaker et al., *Nature Genetics*, 14(4):450-456, 1996 ; Kozal et al., *Nature Medicine*, 2:753-759, 1996). Chips of various formats for use in detecting genetic polymorphisms can be produced on a customized basis by Affymetrix (GeneChip™), Hyseq (HyChip and HyGnostics), and Protogene Laboratories.

In general, these methods employ arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from an individual which, target sequences include a polymorphic marker. EP 785280 describes a tiling strategy for the detection of single nucleotide polymorphisms. Briefly, arrays may generally be “tiling” for a large number of specific polymorphisms. By “tiling” is generally meant the synthesis of a defined set of oligonucleotide probes which is made up of a sequence complementary to the target sequence of interest, as well as preselected variations of that sequence, e.g., substitution of one or more given positions with one or more members of the basis set of monomers, i.e. nucleotides. Tiling strategies are further described in PCT Publication No. WO 95/11995. In a particular aspect, arrays are tiled for a number of specific, identified genetic marker sequences. In particular, the array is tiled to include a number of detection blocks, each detection block being specific for a specific genetic marker or a set of genetic markers. For example, a detection block may be tiled to include a number of probes, which span the sequence segment that includes a specific polymorphism. To ensure probes that are complementary to each allele, the probes are synthesized in pairs differing at the genetic marker. In addition to the probes differing at the polymorphic base, monosubstituted probes are also generally tiled within the detection block. These monosubstituted probes have bases at and up to a certain number of bases in either direction from the polymorphism, substituted with the remaining nucleotides (selected from A, T, G, C and U). Typically the probes in a tiled detection block will include substitutions of the sequence positions up to and including those that are 5 bases away from the genetic marker. The monosubstituted probes provide internal controls for the tiled array, to distinguish actual hybridization from artefactual cross-hybridization. Upon completion of hybridization with the target sequence and washing of the array, the array is scanned to determine the position on the array to which the target sequence hybridizes. The hybridization data from the scanned array is then analyzed

to identify which allele or alleles of the genetic marker are present in the sample. Hybridization and scanning may be carried out as described in PCT Publication No. WO 92/10092 and WO 95/11995 and U.S. Pat. No. 5,424,186.

Thus, in some embodiments, the chips may comprise an array of nucleic acid sequences of fragments of about 15 nucleotides in length. In further embodiments, the chip may comprise an array including at least one of the sequences selected from the group consisting of 9-27, 99-14387, 9-12, 9-13, 99-14405, and 9-16 and the sequences complementary thereto, or a fragment thereof, said fragment comprising at least about 8 consecutive nucleotides, preferably 10, 15, 20, more preferably 25, 30, 40, 47, or 50 consecutive nucleotides and containing a polymorphic base. In preferred embodiments the polymorphic base is within 5, 4, 3, 2, 1, nucleotides of the center of the said polynucleotide, more preferably at the center of said polynucleotide. In some embodiments, the chip may comprise an array of at least 2, 3, 4, 5, 6, 7, 8 or more of these polynucleotides of the invention.

6) Integrated Systems

Another technique, which may be used to analyze polymorphisms, includes multicomponent integrated systems, which miniaturize and compartmentalize processes such as PCR and capillary electrophoresis reactions in a single functional device. An example of such technique is disclosed in U.S. Pat. No. 5,589,136, which describes the integration of PCR amplification and capillary electrophoresis in chips.

Integrated systems can be envisaged mainly when microfluidic systems are used. These systems comprise a pattern of microchannels designed onto a glass, silicon, quartz, or plastic wafer included on a microchip. The movements of the samples are controlled by electric, electroosmotic or hydrostatic forces applied across different areas of the microchip to create functional microscopic valves and pumps with no moving parts. Varying the voltage controls the liquid flow at intersections between the micro-machined channels and changes the liquid flow rate for pumping across different sections of the microchip.

For genotyping genetic markers, the microfluidic system may integrate nucleic acid amplification, microsequencing, capillary electrophoresis and a detection method such as laser-induced fluorescence detection.

In a first step, the DNA samples are amplified, preferably by PCR. Then, the amplification products are subjected to automated microsequencing reactions using ddNTPs (specific fluorescence for each ddNTP) and the appropriate oligonucleotide microsequencing primers which hybridize just upstream of the targeted polymorphic base. Once the extension at the 3' end is completed, the primers are separated from the unincorporated fluorescent ddNTPs by capillary electrophoresis. The separation medium used in capillary electrophoresis can for example be polyacrylamide, polyethyleneglycol or dextran. The incorporated ddNTPs in the single-nucleotide primer extension products are identified by fluorescence detection. This microchip can be used to process at least 96

to 384 samples in parallel. It can use the usual four color laser induced fluorescence detection of the ddNTPs.

7) Allele mapping by MVR-PCR

Minisatellites (VNTR) are composed of tandem repeats 10-100 bp in length, with total array sizes of typically 0.5-50 kb. Polymorphisms exist between tandem repeats generating variant repeat types. The interspersal patterns of variant repeats within alleles can be analyzed by PCR amplification between a universal primer which anneals outside of the repeat array, and primers which binds to specific variant repeats within the array. This technique is called minisatellite variant repeat mapping by PCR, or MVR-PCR. Stead and Jeffreys (2000) *Hum. Mol. Genet.* 9:713-723.

Variant repeat distributions within insulin minisatellite alleles indicate that there are 11 variant repeats (named A-J) based on the 14-bp consensus ACAGGGGTGTGGG (SEQ ID NO:13).

To perform MVR-PCR, insulin minisatellite allele DNA is first prepared. Then, MVR-PCR analysis is performed to determine the fine structure of the allele. In the event that a class III allele is present, it may be necessary to perform reverse MVR-PCR, generating a population of amplification products (amplicons) from the E-repeats to the 3' flanking site. This fine structure analysis allows one to determine the paternal insulin VNTR allele. The procedure is described in more detail in the following paragraphs.

MVR-PCR detects 6 different variant repeats of the insulin minisatellite, the sequences of which are as follows with nucleotides that differ from the A-type repeat consensus underlined:

	<u>Repeat</u>	<u>Sequence</u>	<u>MVR primers</u>
20	A	GTGGGGACAGGGGT (SEQ ID NO:14)	INS-MA
	B	<u>C</u> CTGGGGACAGGGGT (SEQ ID NO:15)	INS-MB and INS-MC
	C	<u>C</u> TGGGGACAGGGGT (SEQ ID NO:16)	INS-MC
	D	<u>C</u> CGGGGACAGGGGT (SEQ ID NO:17)	INS-MD
25	F	<u>C</u> CCGGGGACAGGGGT (SEQ ID NO:18)	INS-MD and INS-MF
	E	GTGGGGATAGGGGT (SEQ ID NO:19)	INS-ME
	H	GTGGGCACAGGGGT (SEQ ID NO:20)	INS-MH

Insulin minisatellite allele DNA is first prepared. Any known method can be used. In general, insulin minisatellite DNA is amplified using PCR primers flanking the minisatellites together with allele-specific primers; amplifying the DNA; separating the alleles on the basis of size, usually on a gel; and extracting the allele DNA from the gel. The following is a non-limiting example. Genomic DNA is amplified by PCR using the following primers: (1) for class I alleles, the forward primer complementary to the flanking site is INS-1296 (5'-ctgctgaggacttgctgcttg-3'; SEQ ID NO:21); and the reverse primer, specific for class I allele is INS-23⁺ (5'-cagaaggacagtgtctgggt-3'; SEQ ID NO:22); and (2) for class III alleles, the forward primer complementary to the flanking site is INS-1296 (SEQ ID NO:21); and the reverse primer, specific for class III allele is INS-23⁻ (5'-cagaaggacagtgtctggga-3'; SEQ ID NO:22). After amplification, PCR products are separated by gel

electrophoresis (e.g., 1% agarose gel); visualized by ethidium bromide staining, and excised from the gel. Class I allele DNA may be released from the gel by adding a dilution buffer, and subjecting the gel to three cycles of freezing/thawing/vortexing. Class III allele DNA may be extracted from the gel using a Qiaex II gel purification kit (Qiagen).

5 MVR-PCR is performed on insulin minisatellite allele DNA. Primers specific for a variant, together with a flanking primer, are used to amplify the allele DNA. Any primer that is specific for a variant can be used. Amplified DNA is subjected to gel electrophoresis, the separated products transferred to a membrane ("blotted"), and the blot analyzed by Southern hybridization using a labeled probe specific for class I allele. The following is a non-limiting example of a suitable
10 protocol. MVR-PCR variant-specific primers are as follows, with a 5' TAG extension indicated in upper case:

INS-MA 5'-TCATGCGTCCATGGTCCGGAaccctgtccccac-3' (SEQ ID NO:23)
INS-MB 5'-TCATGCGTCCATGGTCCGGAaccctgtccccagg-3' (SEQ ID NO:24)
INS-MC 5'-TCATGCGTCCATGGTCCGGAaccctgtccccag-3' (SEQ ID NO:25)
15 INS-MD 5'-TCATGCGTCCATGGTCCGGAaccctgtccccgg-3' (SEQ ID NO:26)
INS-ME 5'-TCATGCGTCCATGGTCCGGAaccctatccccac-3' (SEQ ID NO:27)
INS-MF 5'-TCATGCGTCCATGGTCCGGAaccctgtccccggg-3' (SEQ ID NO:28)
INS-MH 5'-TCATGCGTCCATGGTCCGGAaccctgtccccac-3' (SEQ ID NO:29)

These primers are complementary to the variant-specific sequences A-H, above. 5MVR
20 primers are used together with a flanking site primer (e.g., INS-1296), and, TAG primers. The amplified products are electrophoresed and detected by Southern blot hybridization, as described above.

MVR-PCR of class III alleles accurately types the first approximately 100 repeats in the array. The remainder of the class III allele is typed by creating deletion amplicons covering the 3'
25 end of the array. To achieve this, reverse MVR-PCR is performed using the primers INS-23⁻ and INS-MER, a composite primer with the 3' sequence specific to E-type repeats and the 5' sequence identical to INS-1296. The sequence of INS-MER is 5'-
ctgctgaggacttgctgcttgCAGGGGTGTGGGGAT-3' (SEQ ID NO:30), where the 5' INS-1296
sequence is indicated in lower case. Amplicons thus generated are separated by electrophoresis
30 through a gel, the DNA gel purified, and MVR-PCR mapped as described above. Full allele codes are assembled from overlapping codes generated from the whole allele and each deletion amplicon.

Methods of Genetic Analysis Using Genetic Markers in the INS HphI Locus

Different methods are available for the genetic analysis of complex traits (see Lander and Schork, Science, 265, 2037-2048, 1994). The search for disease-susceptibility genes is conducted
35 using two main methods: the linkage approach in which evidence is sought for cosegregation between a locus and a putative trait locus using family studies, and the association approach in which evidence is sought for a statistically significant association between an allele and a trait or a trait

causing allele (Khoury J. et al., Fundamentals of Genetic Epidemiology, Oxford University Press, NY, 1993). In general, the genetic markers of the present invention find use in any method known in the art to demonstrate a statistically significant correlation between a genotype and a phenotype. The genetic markers may be used in parametric and non-parametric linkage analysis methods. Preferably, the genetic markers of the present invention are used to identify genes associated with detectable traits using association studies, an approach which does not require the use of affected families and which permits the identification of genes associated with complex and sporadic traits.

The genetic analysis using the genetic markers in the INS HphI locuss may be conducted on any scale. The whole set of genetic markers of the present invention or any subset of genetic markers of the present invention corresponding to the candidate gene may be used. Further, any set of genetic markers including a genetic marker of the present invention may be used. A set of genetic polymorphisms that could be used as genetic markers in combination with the genetic markers of the present invention has been described in WO 98/20165. As mentioned above, it should be noted that the genetic markers of the present invention may be included in any complete or partial genetic map of the human genome. These different uses are specifically contemplated in the present invention and claims.

Linkage Analysis

Linkage analysis is based upon establishing a correlation between the transmission of genetic markers and that of a specific trait throughout generations within a family. Thus, the aim of linkage analysis is to detect marker loci that show cosegregation with a trait of interest in pedigrees.

Parametric Methods

When data are available from successive generations there is the opportunity to study the degree of linkage between pairs of loci. Estimates of the recombination fraction enable loci to be ordered and placed onto a genetic map. With loci that are genetic markers, a genetic map can be established, and then the strength of linkage between markers and traits can be calculated and used to indicate the relative positions of markers and genes affecting those (Weir, B.S., Genetic data Analysis II: Methods for Discrete population genetic Data, Sinauer Assoc., Inc., Sunderland, MA, USA, 1996). The classical method for linkage analysis is the logarithm of odds (lod) score method (see Morton N.E., Am.J. Hum.Genet., 7:277-318, 1955; Ott J., Analysis of Human Genetic Linkage, John Hopkins University Press, Baltimore, 1991). Calculation of lod scores requires specification of the mode of inheritance for the disease (parametric method). Generally, the length of the candidate region identified using linkage analysis is between 2 and 20Mb. Once a candidate region is identified as described above, analysis of recombinant individuals using additional markers allows further delineation of the candidate region. Linkage analysis studies have generally relied on the use of a maximum of 5,000 microsatellite markers, thus limiting the maximum theoretical attainable resolution of linkage analysis to about 600 kb on average.

Linkage analysis has been successfully applied to map simple genetic traits that show clear Mendelian inheritance patterns and which have a high penetrance (i.e., the ratio between the number of trait positive carriers of allele *a* and the total number of a carriers in the population). However, parametric linkage analysis suffers from a variety of drawbacks. First, it is limited by its reliance on the choice of a genetic model suitable for each studied trait. Furthermore, as already mentioned, the resolution attainable using linkage analysis is limited, and complementary studies are required to refine the analysis of the typical 2Mb to 20Mb regions initially identified through linkage analysis. In addition, parametric linkage analysis approaches have proven difficult when applied to complex genetic traits, such as those due to the combined action of multiple genes and/or environmental factors. It is very difficult to model these factors adequately in a lod score analysis. In such cases, too large an effort and cost are needed to recruit the adequate number of affected families required for applying linkage analysis to these situations, as recently discussed by Risch, N. and Merikangas, K. (Science, 273:1516-1517, 1996).

Non-Parametric Methods

The advantage of the so-called non-parametric methods for linkage analysis is that they do not require specification of the mode of inheritance for the disease, they tend to be more useful for the analysis of complex traits. In non-parametric methods, one tries to prove that the inheritance pattern of a chromosomal region is not consistent with random Mendelian segregation by showing that affected relatives inherit identical copies of the region more often than expected by chance. Affected relatives should show excess "allele sharing" even in the presence of incomplete penetrance and polygenic inheritance. In non-parametric linkage analysis the degree of agreement at a marker locus in two individuals can be measured either by the number of alleles identical by state (IBS) or by the number of alleles identical by descent (IBD). Affected sib pair analysis is a well-known special case and is the simplest form of these methods.

The genetic markers of the present invention may be used in both parametric and non-parametric linkage analysis. Preferably genetic markers may be used in non-parametric methods which allow the mapping of genes involved in complex traits. The genetic markers of the present invention may be used in both IBD- and IBS- methods to map genes affecting a complex trait. In such studies, taking advantage of the high density of genetic markers, several adjacent genetic marker loci may be pooled to achieve the efficiency attained by multi-allelic markers (Zhao et al., Am. J. Hum. Genet., 63:225-240, (1998).

Population Association Studies

The present invention comprises methods for identifying if the insulin gene or a particular allelic variant thereof is associated with a detectable trait using the genetic markers of the present invention. In one embodiment the present invention comprises methods to detect an association between a genetic marker allele or a genetic marker haplotype and a trait. Further, the invention

comprises methods to identify a trait causing allele in linkage disequilibrium with any genetic marker allele of the present invention.

As described above, alternative approaches can be employed to perform association studies: genome-wide association studies, candidate region association studies and candidate gene association studies. In a preferred embodiment, the genetic markers of the present invention are used to perform candidate gene association studies. The candidate gene analysis clearly provides a short-cut approach to the identification of genes and gene polymorphisms related to a particular trait when some information concerning the biology of the trait is available. Further, the genetic markers of the present invention may be incorporated in any map of genetic markers of the human genome in order to perform genome-wide association studies. Methods to generate a high-density map of genetic markers has been described in PCT Publication No. WO 00/28080. The genetic markers of the present invention may further be incorporated in any map of a specific candidate region of the genome (a specific chromosome or a specific chromosomal segment for example).

As mentioned above, association studies may be conducted within the general population and are not limited to studies performed on related individuals in affected families. Association studies are extremely valuable as they permit the analysis of sporadic or multifactor traits. Moreover, association studies represent a powerful method for fine-scale mapping enabling much finer mapping of trait causing alleles than linkage studies. Studies based on pedigrees often only narrow the location of the trait causing allele. Association studies using the genetic markers of the present invention can therefore be used to refine the location of a trait causing allele in a candidate region identified by Linkage Analysis methods. Moreover, once a chromosome segment of interest has been identified, the presence of a candidate gene, such as a candidate gene of the present invention, in the region of interest can provide a shortcut to the identification of the trait causing allele. Genetic markers of the present invention can be used to demonstrate that a candidate gene is associated with a trait. Such uses are specifically contemplated in the present invention.

Determining the Frequency of a Genetic Marker Allele or of a Genetic Marker Haplotype in a Population

Association studies explore the relationships among frequencies for sets of alleles between loci.

Determining the Frequency of an Allele in a Population

Allelic frequencies of the genetic markers in a populations can be determined using one of the methods described above under the heading "Methods for Genotyping an Individual for Genetic Markers," or any genotyping procedure suitable for this intended purpose. Genotyping pooled samples or individual samples can determine the frequency of a genetic marker allele in a population. One way to reduce the number of genotypings required is to use pooled samples. A major obstacle in using pooled samples is in terms of accuracy and reproducibility for determining accurate DNA concentrations in setting up the pools. Genotyping individual samples provides higher sensitivity,

reproducibility and accuracy and; is the preferred method used in the present invention. Preferably, each individual is genotyped separately and simple gene counting is applied to determine the frequency of an allele of a genetic marker or of a genotype in a given population.

Determining the Frequency of a Haplotype in a Population

5 The gametic phase of haplotypes is unknown when diploid individuals are heterozygous at more than one locus. Using genealogical information in families gametic phase can sometimes be inferred (Perlin et al., *Am. J. Hum. Genet.*, 55:777-787, 1994). When no genealogical information is available different strategies may be used. One possibility is that the multiple-site heterozygous diploids can be eliminated from the analysis, keeping only the homozygotes and the single-site
10 heterozygote individuals, but this approach might lead to a possible bias in the sample composition and the underestimation of low-frequency haplotypes. Another possibility is that single chromosomes can be studied independently, for example, by asymmetric PCR amplification (see Newton et al., *Nucleic Acids Res.*, 17:2503-2516, 1989; Wu et al., *Proc. Natl. Acad. Sci. USA*, 86:2757, 1989), or by isolation of single chromosome by limit dilution followed by PCR
15 amplification (see Ruano et al., *Proc. Natl. Acad. Sci. USA*, 87:6296-6300, 1990). Further, a sample may be haplotyped for sufficiently close genetic markers by double PCR amplification of specific alleles (Sarkar, G. and Sommer S.S., *Biotechniques*, 1991). These approaches are not entirely satisfying either because of their technical complexity, the additional cost they entail, their lack of generalisation at a large scale, or the possible biases they introduce. To overcome these difficulties,
20 an algorithm to infer the phase of PCR-amplified DNA genotypes introduced by Clark A.G. (*Mol. Biol. Evol.*, 7:111-122, 1990), may be used. Briefly, the principle is to start filling a preliminary list of haplotypes present in the sample by examining unambiguous individuals, that is, the complete homozygotes and the single-site heterozygotes. Then other individuals in the same sample are screened for the possible occurrence of previously recognised haplotypes. For each positive
25 identification, the complementary haplotype is added to the list of recognised haplotypes, until the phase information for all individuals is either resolved or identified as unresolved. This method assigns a single haplotype to each multiheterozygous individual, whereas several haplotypes are possible when there are more than one heterozygous site. Alternatively, one can use methods estimating haplotype frequencies in a population without assigning haplotypes to each individual.
30 Preferably, a method based on an expectation-maximization (EM) algorithm (Dempster et al., *J. R. Stat. Soc.*, 39B: 1-38, 1977), leading to maximum-likelihood estimates of haplotype frequencies under the assumption of Hardy-Weinberg proportions (random mating) is used (see Excoffier L. and Slatkin M., *Mol. Biol. Evol.*, 12(5): 921-927, 1995). The EM algorithm is a generalised iterative maximum-likelihood approach to estimation that is useful when data are ambiguous and/or
35 incomplete. The EM algorithm is used to resolve heterozygotes into haplotypes. Haplotype estimations are further described below under the heading "Statistical methods." Any other method

known in the art to determine or to estimate the frequency of a haplotype in a population may also be used.

Linkage Disequilibrium Analysis

Linkage disequilibrium is the non-random association of alleles at two or more loci and represents a powerful tool for mapping genes involved in disease traits (see Ajioka R.S. et al., *Am. J. Hum. Genet.*, 60:1439-1447, 1997). Genetic markers, because they are densely spaced in the human genome and can be genotyped in greater numbers than other types of genetic markers (such as RFLP or VNTR markers), are particularly useful in genetic analysis based on linkage disequilibrium.

When a disease mutation is first introduced into a population (by a new mutation or the immigration of a mutation carrier), it necessarily resides on a single chromosome and thus on a single “background” or “ancestral” haplotype of linked markers. Consequently, there is complete disequilibrium between these markers and the disease mutation: one finds the disease mutation only in the presence of a specific set of marker alleles. Through subsequent generations recombination events occur between the disease mutation and these marker polymorphisms, and the disequilibrium gradually dissipates. The pace of this dissipation is a function of the recombination frequency, so the markers closest to the disease gene will manifest higher levels of disequilibrium than those that are further away. When not broken up by recombination, “ancestral” haplotypes and linkage disequilibrium between marker alleles at different loci can be tracked not only through pedigrees but also through populations. Linkage disequilibrium is usually seen as an association between one specific allele at one locus and another specific allele at a second locus.

The pattern or curve of disequilibrium between disease and marker loci is expected to exhibit a maximum that occurs at the disease locus. Consequently, the amount of linkage disequilibrium between a disease allele and closely linked genetic markers may yield valuable information regarding the location of the disease gene. For fine-scale mapping of a disease locus, it is useful to have some knowledge of the patterns of linkage disequilibrium that exist between markers in the studied region. As mentioned above the mapping resolution achieved through the analysis of linkage disequilibrium is much higher than that of linkage studies. The high density of genetic markers combined with linkage disequilibrium analysis provides powerful tools for fine-scale mapping. Different methods to calculate linkage disequilibrium are described below under the heading “Statistical Methods.”

Population-Based Case-Control Studies of Trait-Marker Associations

As mentioned above, the occurrence of pairs of specific alleles at different loci on the same chromosome is not random and the deviation from random is called linkage disequilibrium. Association studies focus on population frequencies and rely on the phenomenon of linkage disequilibrium. If a specific allele in a given gene is directly involved in causing a particular trait, its frequency will be statistically increased in an affected (trait positive) population, when compared to the frequency in a trait negative population or in a random control population. As a consequence of the existence of linkage disequilibrium, the frequency of all other alleles present in the haplotype

carrying the trait-causing allele will also be increased in trait positive individuals compared to trait negative individuals or random controls. Therefore, association between the trait and any allele (specifically a genetic marker allele) in linkage disequilibrium with the trait-causing allele will suffice to suggest the presence of a trait-related gene in that particular region. Case-control

5 populations can be genotyped for genetic markers to identify associations that narrowly locate a trait causing allele. As any marker in linkage disequilibrium with one given marker associated with a trait will be associated with the trait. Linkage disequilibrium allows the relative frequencies in case-control populations of a limited number of genetic polymorphisms (specifically genetic markers) to be analyzed as an alternative to screening all possible functional polymorphisms in order to find trait-

10 causing alleles. Association studies compare the frequency of marker alleles in unrelated case-control populations, and represent powerful tools for the dissection of complex traits.

Case-Control Populations (Inclusion Criteria)

Population-based association studies do not concern familial inheritance but compare the prevalence of a particular genetic marker, or a set of markers, in case-control populations. They are

15 case-control studies based on comparison of unrelated case (affected or trait positive) individuals and unrelated control (unaffected, trait negative or random) individuals. Preferably the control group is composed of unaffected or trait negative individuals. Further, the control group is ethnically matched to the case population. Moreover, the control group is preferably matched to the case-population for the main known confusion factor for the trait under study (for example age-matched for an age-

20 dependent trait). Ideally, individuals in the two samples are paired in such a way that they are expected to differ only in their disease status. The terms "trait positive population," "case population" and "affected population" are used interchangeably herein.

An important step in the dissection of complex traits using association studies is the choice of case-control populations (see Lander and Schork, Science, 265, 2037-2048, 1994). A major step

25 in the choice of case-control populations is the clinical definition of a given trait or phenotype. Any genetic trait may be analyzed by the association method proposed here by carefully selecting the individuals to be included in the trait positive and trait negative phenotypic groups. Four criteria are often useful: clinical phenotype, age at onset, family history and severity. The selection procedure for continuous or quantitative traits (such as blood pressure for example) involves selecting

30 individuals at opposite ends of the phenotype distribution of the trait under study, so as to include in these trait positive and trait negative populations individuals with non-overlapping phenotypes. Preferably, case-control populations consist of phenotypically homogeneous populations. Trait positive and trait negative populations consist of phenotypically uniform populations of individuals representing each between 1 and 98%, preferably between 1 and 80%, more preferably between 1

35 and 50%, and more preferably between 1 and 30%, most preferably between 1 and 20% of the total population under study, and preferably selected among individuals exhibiting non-overlapping phenotypes. The clearer the difference between the two trait phenotypes, the greater the probability

of detecting an association with genetic markers. The selection of those drastically different but relatively uniform phenotypes enables efficient comparisons in association studies and the possible detection of marked differences at the genetic level, provided that the sample sizes of the populations under study are significant enough.

5 In preferred embodiments, a first group of between 50 and 300 trait positive individuals, preferably about 100 individuals, are recruited according to their phenotypes. A similar number of trait negative individuals are included in such studies.

 In the present invention, typical examples of inclusion criteria include obesity, diabetic, ethnicity, monotonic gain of weight, age, gender and puberty.

10 Suitable examples of association studies using genetic markers including the genetic markers of the present invention, are studies involving the following populations: (1) a case population suffering from juvenile onset obesity and a lean control population; and (2) an adult case population suffering from obesity and an age-matched lean control population.

 In an embodiment, markers in linkage disequilibrium with the insulin HphI locus may be
15 used to identify individuals who are prone to obesity. This includes diagnostic and prognostic assays to identify individuals who possess factors which predispose them to obesity, as well as clinical trials and treatment regimens which utilize these assays. Drug treatment may include any pharmaceutical compound suspected or known in the art used to treat obesity or control obesity, and disorders associated with obesity.

20 Association Analysis

 The general strategy to perform association studies using genetic markers derived from a region carrying a candidate gene is to scan two groups of individuals (case-control populations) in order to measure and statistically compare the allele frequencies of the genetic markers of the present invention in both groups.

25 If a statistically significant association with a trait is identified for at least one or more of the analyzed genetic markers, one can assume that: either the associated allele is directly responsible for causing the trait (i.e. the associated allele is the trait causing allele), or more likely the associated allele is in linkage disequilibrium with the trait causing allele. The specific characteristics of the associated allele with respect to the candidate gene function usually give further insight into the
30 relationship between the associated allele and the trait (causal or in linkage disequilibrium). If the evidence indicates that the associated allele within the candidate gene is most probably not the trait causing allele but is in linkage disequilibrium with the real trait causing allele, then the trait causing allele can be found by sequencing the vicinity of the associated marker, and performing further association studies with the polymorphisms that are revealed in an iterative manner.

35 Association studies are usually run in two successive steps. In a first phase, the frequencies of a reduced number of genetic markers from the candidate gene are determined in the trait positive and trait negative populations. In a second phase of the analysis, the position of the genetic loci

responsible for the given trait is further refined using a higher density of markers from the relevant region.

Haplotype Analysis

As described above, when a chromosome carrying a disease allele first appears in a population as a result of either mutation or migration, the mutant allele necessarily resides on a chromosome having a set of linked markers: the ancestral haplotype. This haplotype can be tracked through populations and its statistical association with a given trait can be analyzed. Complementing single point (allelic) association studies with multi-point association studies also called haplotype studies increases the statistical power of association studies. Thus, a haplotype association study allows one to define the frequency and the type of the ancestral carrier haplotype. A haplotype analysis is important in that it increases the statistical power of an analysis involving individual markers.

In a first stage of a haplotype frequency analysis, the frequency of the possible haplotypes based on various combinations of the identified genetic markers of the invention is determined. The haplotype frequency is then compared for distinct populations of trait positive and control individuals. The number of trait positive individuals, which should be, subjected to this analysis to obtain statistically significant results usually ranges between 30 and 300, with a preferred number of individuals ranging between 50 and 150. The same considerations apply to the number of unaffected individuals (or random control) used in the study. The results of this first analysis provide haplotype frequencies in case-control populations, for each evaluated haplotype frequency a p-value and an odd ratio are calculated. If a statistically significant association is found the relative risk for an individual carrying the given haplotype of being affected with the trait under study can be approximated.

Interaction Analysis

Genetic markers described above may also be used to identify patterns of genetic markers associated with detectable traits resulting from polygenic interactions. The analysis of genetic interaction between alleles at unlinked loci requires individual genotyping using the techniques described herein. The analysis of allelic interaction among a selected set of genetic markers with appropriate level of statistical significance can be considered as a haplotype analysis. Interaction analysis consists in stratifying the case-control populations with respect to a given haplotype for the first loci and performing a haplotype analysis with the second loci with each subpopulation.

Testing For Linkage in the Presence of Association

The genetic markers described above may further be used in TDT (transmission/disequilibrium test). TDT tests for both linkage and association and is not affected by population stratification. TDT requires data for affected individuals and their parents or data from unaffected sibs instead of from parents (see Spielmann S. et al., 1993; Schaid D.J. et al., 1996, Spielmann S. and Ewens W.J., 1998). Such combined tests generally reduce the false – positive errors produced by separate analyses.

Statistical Methods

In general, any method known in the art to test whether a trait and a genotype show a statistically significant correlation may be used.

1) Methods In Linkage Analysis

5 Statistical methods and computer programs useful for linkage analysis are well-known to those skilled in the art (see Terwilliger J.D. and Ott J., Handbook of Human Genetic Linkage, John Hopkins University Press, London, 1994; Ott J., Analysis of Human Genetic Linkage, John Hopkins University Press, Baltimore, 1991).

2) Methods to Estimate Haplotype Frequencies in a Population

10 As described above, when genotypes are scored, it is often not possible to distinguish heterozygotes so that haplotype frequencies cannot be easily inferred. When the gametic phase is not known, haplotype frequencies can be estimated from the multilocus genotypic data. Any method known to person skilled in the art can be used to estimate haplotype frequencies (see Lange K., Mathematical and Statistical Methods for Genetic Analysis, Springer, New York, 1997; Weir, B.S., Genetic data Analysis II: Methods for Discrete population genetic Data, Sinauer Assoc., Inc., Sunderland, MA, USA, 1996). Preferably, maximum-likelihood haplotype frequencies are computed using an Expectation- Maximization (EM) algorithm (see Dempster et al., J. R. Stat. Soc., 39B:1-38, 1977; Excoffier L. and Slatkin M., Mol. Biol. Evol., 12(5): 921-927, 1995). This procedure is an iterative process aiming at obtaining maximum-likelihood estimates of haplotype frequencies from multi-locus genotype data when the gametic phase is unknown. Haplotype estimations are usually performed by applying the EM algorithm using for example the EM-HAPLO program (Hawley M.E. et al., Am. J. Phys. Anthropol., 18:104, 1994) or the Arlequin program (Schneider et al., Arlequin: a software for population genetics data analysis, University of Geneva, 1997). The EM algorithm is a generalised iterative maximum likelihood approach to estimation and is briefly described below.

25 In what follows, phenotypes will refer to multi-locus genotypes with unknown haplotypic phase. Genotypes will refer to mutli-locus genotypes with known haplotypic phase.

Suppose one has a sample of N unrelated individuals typed for K markers. The data observed are the unknown-phase K-locus phenotypes that can be categorized with F different phenotypes. Further, suppose that we have H possible haplotypes (in the case of K genetic markers, we have for the maximum number of possible haplotypes $H=2^K$).

For phenotype j with c_j possible genotypes, we have:

$$P_j = \sum_{i=1}^{c_j} P(\text{genotype}(i)) = \sum_{i=1}^{c_j} P(h_k, h_l). \quad \text{Equation 1}$$

Here, P_j is the probability of the j^{th} phenotype, and $P(h_k, h_l)$ is the probability of the i^{th} genotype composed of haplotypes h_k and h_l . Under random mating (i.e. Hardy-Weinberg Equilibrium), $P(h_k h_l)$ is expressed as:

$$P(h_k, h_l) = P(h_k)^2 \text{ for } h_k = h_l, \text{ and}$$

$$P(h_k, h_l) = 2P(h_k)P(h_l) \text{ for } h_k \neq h_l. \quad \text{Equation 2}$$

The E-M algorithm is composed of the following steps: First, the genotype frequencies are estimated from a set of initial values of haplotype frequencies. These haplotype frequencies are denoted $P_1^{(0)}, P_2^{(0)}, P_3^{(0)}, \dots, P_H^{(0)}$. The initial values for the haplotype frequencies may be obtained from a random number generator or in some other way well known in the art. This step is referred to the Expectation step. The next step in the method, called the Maximization step, consists of using the estimates for the genotype frequencies to re-calculate the haplotype frequencies. The first iteration haplotype frequency estimates are denoted by $P_1^{(1)}, P_2^{(1)}, P_3^{(1)}, \dots, P_H^{(1)}$. In general, the Expectation step at the s^{th} iteration consists of calculating the probability of placing each phenotype into the different possible genotypes based on the haplotype frequencies of the previous iteration:

$$P(h_k, h_l)^{(s)} = \frac{n_j}{N} \left[\frac{P_j(h_k, h_l)^{(s)}}{P_j} \right], \quad \text{Equation 3}$$

where n_j is the number of individuals with the j^{th} phenotype and $P_j(h_k, h_l)^{(s)}$ is the probability of genotype h_k, h_l in phenotype j . In the Maximization step, which is equivalent to the gene-counting method (Smith, Ann. Hum. Genet., 21:254-276, 1957), the haplotype frequencies are re-estimated based on the genotype estimates:

$$P_t^{(s+1)} = \frac{1}{2} \sum_{j=1}^F \sum_{i=1}^{c_j} \delta_{it} P_j(h_k, h_l)^{(s)}. \quad \text{Equation 4}$$

Here, δ_{it} is an indicator variable which counts the number of occurrences that haplotype t is present in i^{th} genotype; it takes on values 0, 1, and 2.

The E-M iterations cease when the following criterion has been reached. Using Maximum Likelihood Estimation (MLE) theory, one assumes that the phenotypes j are distributed multinomially. At each iteration s , one can compute the likelihood function L . Convergence is achieved when the difference of the log-likelihood between two consecutive iterations is less than some small number, preferably 10^{-7} .

3) Methods To Calculate Linkage Disequilibrium Between Markers

A number of methods can be used to calculate linkage disequilibrium between any two genetic positions, in practice linkage disequilibrium is measured by applying a statistical association test to haplotype data taken from a population.

Linkage disequilibrium between any pair of genetic markers comprising at least one of the genetic markers of the present invention (M_i, M_j) having alleles (a_i/b_i) at marker M_i and alleles (a_j/b_j) at marker M_j can be calculated for every allele combination ($a_i, a_j; a_i, b_j; b_i, a_j$ and b_i, b_j), according to the Piazza formula :

$$\Delta_{ai aj} = \sqrt{\theta_4} - \sqrt{(\theta_4 + \theta_3)(\theta_4 + \theta_2)}, \text{ where :}$$

$\theta_4 = - -$ = frequency of genotypes not having allele a_i at M_i and not having allele a_j at M_j

$\theta_3 = - +$ = frequency of genotypes not having allele a_i at M_i and having allele a_j at M_j

$\theta_2 = + -$ = frequency of genotypes having allele a_i at M_i and not having allele a_j at M_j

Linkage disequilibrium (LD) between pairs of genetic markers (M_i , M_j) can also be calculated for every allele combination (a_i, a_j , a_i, b_j , b_i, a_j and b_i, b_j), according to the maximum-likelihood estimate (MLE) for delta (the composite genotypic disequilibrium coefficient), as described by Weir (Weir B. S., 1996). The MLE for the composite linkage disequilibrium is:

$$D_{aiaj} = (2n_1 + n_2 + n_3 + n_4/2)/N - 2(\text{pr}(a_i) \cdot \text{pr}(a_j))$$

Where $n_1 = \Sigma$ phenotype (a_i/a_i , a_j/a_j), $n_2 = \Sigma$ phenotype (a_i/a_i , a_j/b_j), $n_3 = \Sigma$ phenotype (a_i/b_i , a_j/a_j), $n_4 = \Sigma$ phenotype (a_i/b_i , a_j/b_j) and N is the number of individuals in the sample.

This formula allows linkage disequilibrium between alleles to be estimated when only genotype, and not haplotype, data are available.

Another means of calculating the linkage disequilibrium between markers is as follows. For a couple of genetic markers, M_i (a_i/b_i) and M_j (a_j/b_j), fitting the Hardy-Weinberg equilibrium, one can estimate the four possible haplotype frequencies in a given population according to the approach described above.

The estimation of gametic disequilibrium between a_i and a_j is simply:

$$D_{aiaj} = \text{pr}(\text{haplotype}(a_i, a_j)) - \text{pr}(a_i) \cdot \text{pr}(a_j).$$

Where $\text{pr}(a_i)$ is the probability of allele a_i and $\text{pr}(a_j)$ is the probability of allele a_j and where $\text{pr}(\text{haplotype}(a_i, a_j))$ is estimated as in Equation 3 above.

For a couple of genetic markers only one measure of disequilibrium is necessary to describe the association between M_i and M_j .

Then a normalized value of the above is calculated as follows:

$$D'_{aiaj} = D_{aiaj} / \max(-\text{pr}(a_i) \cdot \text{pr}(a_j), -\text{pr}(b_i) \cdot \text{pr}(b_j)) \text{ with } D_{aiaj} < 0$$

$$D'_{aiaj} = D_{aiaj} / \max(\text{pr}(b_i) \cdot \text{pr}(a_j), \text{pr}(a_i) \cdot \text{pr}(b_j)) \text{ with } D_{aiaj} > 0$$

The skilled person will readily appreciate that other LD calculation methods can be used.

Linkage disequilibrium among a set of genetic markers having an adequate heterozygosity rate can be determined by genotyping between 50 and 1000 unrelated individuals, preferably between 75 and 200, more preferably around 100.

4) Testing For Association

Methods for determining the statistical significance of a correlation between a phenotype and a genotype, in this case an allele at a genetic marker or a haplotype made up of such alleles, may be determined by any statistical test known in the art and with any accepted threshold of statistical significance being required. The application of particular methods and thresholds of significance are well with in the skill of the ordinary practitioner of the art.

Testing for association is performed by determining the frequency of a genetic marker allele in case and control populations and comparing these frequencies with a statistical test to determine if there is a statistically significant difference in frequency which would indicate a correlation between the trait and the genetic marker allele under study. Similarly, a haplotype analysis is performed by

estimating the frequencies of all possible haplotypes for a given set of genetic markers in case and control populations, and comparing these frequencies with a statistical test to determine if there is a statistically significant correlation between the haplotype and the phenotype (trait) under study. Any statistical tool useful to test for a statistically significant association between a genotype and a phenotype may be used. Preferably the statistical test employed is a chi-square test with one degree of freedom. A P-value is calculated (the P-value is the probability that a statistic as large or larger than the observed one would occur by chance).

Statistical Significance

In preferred embodiments, significance for diagnosis purposes, either as a positive basis for further diagnostic tests or as a preliminary starting point for early preventive therapy, the p value related to a genetic marker association is preferably about 1×10^{-2} or less, more preferably about 1×10^{-4} or less, for a single genetic marker analysis and about 1×10^{-3} or less, still more preferably 1×10^{-6} or less and most preferably of about 1×10^{-8} or less, for a haplotype analysis involving two or more markers. These values are believed to be applicable to any association studies involving single or multiple marker combinations.

The skilled person can use the range of values set forth above as a starting point in order to carry out association studies with genetic markers of the present invention. In doing so, significant associations between the genetic markers of the present invention and obesity or disorders related to obesity can be revealed and used for diagnosis and drug screening purposes.

Phenotypic Permutation

In order to confirm the statistical significance of the first stage haplotype analysis described above, it might be suitable to perform further analyses in which genotyping data from case-control individuals are pooled and randomized with respect to the trait phenotype. Each individual genotyping data is randomly allocated to two groups, which contain the same number of individuals as the case-control populations used to compile the data obtained in the first stage. A second stage haplotype analysis is preferably run on these artificial groups, preferably for the markers included in the haplotype of the first stage analysis showing the highest relative risk coefficient. This experiment is reiterated preferably at least between 100 and 10000 times. The repeated iterations allow the determination of the percentage of obtained haplotypes with a significant p-value level below about 1×10^{-3} .

Assessment of Statistical Association

To address the problem of false positives similar analysis may be performed with the same case-control populations in random genomic regions. Results in random regions and the candidate region are compared as described in PCT Publication No. WO 00/28080.

5 5) Evaluation of Risk Factors

The association between a risk factor (in genetic epidemiology the risk factor is the presence or the absence of a certain allele or haplotype at marker loci) and a disease is measured by the odds ratio (OR) and by the relative risk (RR). If $P(R^+)$ is the probability of developing the disease for individuals with R and $P(R^-)$ is the probability for individuals without the risk factor, then the relative risk is simply the ratio of the two probabilities, that is:

$$RR = P(R^+)/P(R^-)$$

In case-control studies, direct measures of the relative risk cannot be obtained because of the sampling design. However, the odds ratio allows a good approximation of the relative risk for low-incidence diseases and can be calculated:

$$15 \quad OR = (F^+/(1-F^+))/(F^-/(1-F^-))$$

F^+ is the frequency of the exposure to the risk factor in cases and F^- is the frequency of the exposure to the risk factor in controls. F^+ and F^- are calculated using the allelic or haplotype frequencies of the study and further depend on the underlying genetic model (dominant, recessive, additive, etc).

20 One can further estimate the attributable risk (AR) which describes the proportion of individuals in a population exhibiting a trait due to a given risk factor. This measure is important in quantifying the role of a specific factor in disease etiology and in terms of the public health impact of a risk factor. The public health relevance of this measure lies in estimating the proportion of cases of disease in the population that could be prevented if the exposure of interest were absent. AR is determined as follows:

$$AR = P_E (RR-1) / (P_E (RR-1)+1)$$

AR is the risk attributable to a genetic marker allele or a genetic marker haplotype. P_E is the frequency of exposure to an allele or a haplotype within the population at large; and RR is the relative risk which, is approximated with the odds ratio when the trait under study has a relatively low incidence in the general population.

Identification of Genetic markers in Linkage Disequilibrium with the Genetic markers of the Invention

Once a first genetic marker has been identified in a genomic region of interest, a practitioner of ordinary skill in the art, using the teachings of the present invention, can easily identify additional genetic markers in linkage disequilibrium with this first marker. As mentioned before any marker in linkage disequilibrium with a first marker associated with a trait will be associated with the trait. Therefore, once an association has been demonstrated between a given genetic marker and a trait, the

discovery of additional genetic markers associated with this trait is of great interest in order to increase the density of genetic markers in this particular region. The causal gene or mutation will be found in the vicinity of the marker or set of markers showing the highest correlation with the trait.

Identification of additional markers in linkage disequilibrium with a given marker involves:

- 5 (a) amplifying a genomic fragment comprising a first genetic marker from a plurality of individuals;
- (b) identifying of second genetic markers in the genomic region harboring the first genetic marker;
- (c) conducting a linkage disequilibrium analysis between the first genetic marker and second genetic markers; and (d) selecting the second genetic markers as being in linkage disequilibrium with the first marker. Subcombinations comprising steps (b) and (c) are also contemplated.

10 Methods to identify genetic markers and to conduct linkage disequilibrium analysis are described herein and can be carried out by the skilled person without undue experimentation. Genetic markers which are in linkage disequilibrium with the insulin HphI locus, which are expected to present similar characteristics in terms of their respective association with a given trait, e.g. obesity, can be used. The HphI locus is in strong linkage disequilibrium with the neighboring insulin VNTR: the '+' alleles (T) of the HphI locus are in complete linkage disequilibrium with class I alleles of the neighboring insulin VNTR, and '-' alleles (A) with the class III alleles. Therefore, linkage disequilibrium analysis also tests the insulin VNTR through the -23 HphI polymorphism as a surrogate marker. Optionally, wherein the marker in linkage disequilibrium with the insulin HphI locus is selected from the group consisting of markers described in Table C; preferably markers -
15 4217 PstI, -2221 MspI, -23 HphI, +1428 FokI, +11000 AluI and +32000 ApaI; or more preferably marker -23 HphI. Optionally, the marker in linkage disequilibrium with the insulin HphI locus may further include any other marker that is in linkage disequilibrium with the insulin HphI locus that is known in the art; as well as any marker determined to be in linkage disequilibrium with the insulin HphI locus by methods described herein.

25 Mapping Studies: Identification of Functional Mutations

Once a positive association is confirmed with a genetic marker of the present invention, sequence in the associated candidate region (within linkage disequilibrium of the insulin gene) can be scanned for mutations by comparing the sequences of a selected number of trait positive and trait negative individuals. In a preferred embodiment, functional regions such as exons and splice sites, promoters and other regulatory regions of the insulin gene are scanned for mutations. Preferably, trait positive individuals carry the haplotype shown to be associated with the trait, and trait negative individuals do not carry the haplotype or allele associated with the trait. The mutation detection procedure is essentially similar to that used for biallelic site identification.

The method used to detect such mutations generally comprises the following steps: (a)
35 amplification of a region of the candidate gene comprising a genetic marker or a group of genetic markers associated with the trait from DNA samples of trait positive patients and trait negative controls; (b) sequencing of the amplified region; (c) comparison of DNA sequences from trait-

positive patients and trait-negative controls; and (d) determination of mutations specific to trait-positive patients. Subcombinations which comprise steps (b) and (c) are specifically contemplated.

It is preferred that candidate polymorphisms be then verified by screening a larger population of cases and controls by means of any genotyping procedure such as those described herein, preferably using a microsequencing technique in an individual test format. Polymorphisms are considered as candidate mutations when present in cases and controls at frequencies compatible with the expected association results.

Genetic markers of the Invention in Methods of Genetic Diagnostics

The genetic markers of the present invention can also be used to develop diagnostic tests capable of identifying individuals who express a detectable trait as the result of a specific genotype or individuals whose genotype places them at risk of developing a detectable trait at a subsequent time.

It will of course be understood by practitioners skilled in the treatment or diagnosis of obesity and disorders related to obesity that the present invention does not intend to provide an absolute identification of individuals who could be at risk of developing a particular disease involving obesity and disorders related to obesity but rather to indicate a certain degree or likelihood of developing a disease. However, this information is extremely valuable as it can, in certain circumstances, be used to initiate preventive treatments or to allow an individual carrying a significant haplotype to foresee warning signs such as minor symptoms. In diseases in which attacks may be extremely severe and sometimes fatal if not treated on time, the knowledge of a potential predisposition, even if this predisposition is not absolute, might contribute in a very significant manner to treatment efficacy.

The diagnostic techniques of the present invention may employ a variety of methodologies to determine whether a test subject has a genetic marker pattern associated with an increased risk of developing a detectable trait or whether the individual suffers from a detectable trait as a result of a particular mutation, including methods which enable the analysis of individual chromosomes for haplotyping, such as family studies, single sperm DNA analysis or somatic hybrids. The trait analyzed using the present diagnostics may be any detectable trait, including obesity and disorders related to obesity.

Another aspect of the present invention relates to a method of determining whether an individual is at risk of developing a trait or whether an individual expresses a trait as a consequence of possessing a particular trait-causing allele. The present invention also relates to a method of determining whether an individual is at risk of developing a plurality of traits or whether an individual expresses a plurality of traits as a result of possessing a particular trait-causing allele. These methods involve obtaining a nucleic acid sample from the individual and determining whether the nucleic acid sample contains one or more alleles of one or more genetic markers indicative of a risk of developing the trait or indicative that the individual expresses the trait as a result of possessing a particular trait-causing allele. These methods also involve obtaining a nucleic acid sample from

the individual and, determining, whether the nucleic acid sample contains at least one allele or at least one genetic marker haplotype, indicative of a risk of developing the trait or indicative that the individual expresses the trait as a result of possessing a particular insulin polymorphism or mutation (trait-causing allele).

5 Preferably, in such diagnostic methods, a nucleic acid sample is obtained from the individual and this sample is genotyped using methods described above in "Methods Of Genotyping DNA Samples For Genetic Markers." The diagnostics may be based on a single genetic marker or on a group of genetic markers. In each of these methods, a nucleic acid sample is obtained from the test subject and the genetic marker pattern of one or more of the markers in linkage disequilibrium with
10 the insulin HphI locus is determined. Alternatively, the one or more genetic markers are selected from the group of markers described in Table C; preferably markers -4217 PstI, -2221 MspI, -23 HphI, +1428 FokI, +11000 AluI and +32000 ApaI; or more preferably marker -23 HphI. Optionally, the marker in linkage disequilibrium with the insulin HphI locus may further include any other marker that is in linkage disequilibrium with the insulin HphI locus that is known in the art; as well as any
15 marker determined to be in linkage disequilibrium with the insulin HphI locus by methods described herein.

 In one embodiment, a PCR amplification is conducted on the nucleic acid sample to amplify regions in which polymorphisms associated with a detectable phenotype have been identified. The amplification products are sequenced to determine whether the individual possesses one or more
20 insulin polymorphisms associated with a detectable phenotype. The primers used to generate amplification products may comprise the primers listed in Table C and Table Amplification Primers. Alternatively, the nucleic acid sample is subjected to microsequencing reactions as described above to determine whether the individual possesses one or more insulin polymorphisms associated with a detectable phenotype resulting from a mutation or a polymorphism in the insulin gene.

25 In another embodiment, the nucleic acid sample is contacted with one or more allele specific oligonucleotide probes which specifically hybridize to one or more insulin alleles associated with a detectable phenotype. In another embodiment, the nucleic acid sample is contacted with a second insulin oligonucleotide capable of producing an amplification product when used with the allele specific oligonucleotide in an amplification reaction. The presence of an amplification product in the
30 amplification reaction indicates that the individual possesses one or more insulin-related alleles associated with a detectable phenotype.

 As described herein, the diagnostics may be based on a single genetic marker or a group of genetic markers. Preferably, the genetic marker or combination of genetic markers is selected from the group consisting of markers in linkage disequilibrium with the insulin HphI locus described in Table
35 A; preferably markers -4217 PstI, -2221 MspI, -23 HphI, +1428 FokI, +11000 AluI and +32000 ApaI; or more preferably marker -23 HphI. Optionally, the marker in linkage disequilibrium with the insulin HphI locus may further include any other marker that is in linkage disequilibrium with the

insulin HphI locus that is known in the art; as well as any marker determined to be in linkage disequilibrium with the insulin HphI locus by methods described herein. Diagnostic kits may comprise any of the polynucleotides of the present invention.

These diagnostic methods are extremely valuable as they can, in certain circumstances, be used to initiate preventive treatments or to allow an individual carrying a significant genotype or haplotype to foresee warning signs such as minor symptoms. For example, in the study described in Example 1, the subjects were all obese juveniles. However, by identifying infants or toddlers who carry a paternal VNTR Class I allele, infants and toddlers who are at risk for becoming obese, such individuals could be targeted now for modulation of dietary intake of calories to prevent the onset of later severe disease.

Diagnostics, which analyze and predict response to a drug or side effects to a drug, may be used to determine whether an individual should be treated with a particular drug. For example, if the diagnostic indicates a likelihood that an individual will respond positively to treatment with a particular drug, the drug may be administered to the individual. Conversely, if the diagnostic indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects. Other associations between markers in linkage disequilibrium with the insulin HphI locus and other traits associated with insulin-related disorders can also be determined using the methods of the invention without undue experimentation and would indicate other markers useful to identify sub-populations of people likely to be susceptible (or not) to a drug targeting those traits. In addition, specific associations can be performed looking at drug outcome (treatment/side effect) to identify other useful markers for predicting risks/successful treatment.

Clinical drug trials represent another application for the markers of the present invention. One or more markers indicative of response to an agent acting against an insulin-related disorder or to side effects to an agent acting against an insulin-related disorder may be identified using the methods described above. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and/or exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who have the potential to respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and/or without risking undesirable safety problems.

TREATMENT OF OBESITY

The invention further provides methods of treating obesity, e.g., prophylactic methods of treating obesity. The invention further provides methods of treating, e.g., prophylactic methods, disorders related to obesity. The methods generally comprise determining the INS VNTR genotype of an individual, as described above; and, where the individual has a paternal VNTR Class I allele,

submitting the individual to a weight control regimen. In some embodiments, the invention provides methods for reducing the risk that an individual will develop obesity. In some of these embodiments, the obesity is early onset obesity. In other embodiments, the invention provides methods for reducing the risk that an individual will develop a disorder related to obesity.

5 The proposed treatments for reducing body weight (controlling body weight) are of five types. (1) Food restriction is the most frequently used. The obese individuals are advised to change their dietary habits so as to consume fewer calories, i.e. a very low calorie (VLC) diet (400 and 800 kcal/day). Although this type of treatment is effective in the short-term, the recidivation rate is very high. (2) Increased calorie use through physical exercise is also proposed. This treatment is
10 ineffective when applied alone, but it improves weight-loss in subjects on a low-calorie diet. Together, food restriction and increased calorie use are sometimes considered a single behavioral modification treatment. (3) Gastrointestinal surgery, which reduces the absorption of the calories ingested, is effective, but has been virtually abandoned because of the side effects it causes. (4) An approach that aims to reduce the absorption of dietary lipids by sequestering them in the lumen of the
15 digestive tube is also in place. However, it induces physiological imbalances which are difficult to tolerate, including: deficiency in the absorption of fat-soluble vitamins, flatulence and steatorrhea. Whatever the envisaged therapeutic approach, the treatments of obesity are all characterized by an extremely high recidivation rate. (5) There are five medicinal strategies that may lead to significant weight loss:

20 reducing food intake by amplifying inhibitory effects of anorexigenic signals or factors (those that suppress food intake) or by blocking orexigenic signals or factors (those that stimulate food intake), i.e. sibutramine;
 blocking nutrient absorption (especially fat) in the gut, i.e. orlistat;
 increasing thermogenesis by uncoupling of fuel metabolism from the generation of ATP,
25 thereby dissipating food energy as heat, i.e. ephedrine and caffeine;
 modulating fat or protein metabolism or storage by regulating fat synthesis/lipolysis or adipose differentiation/apoptosis; and
 modulating the central controller regulating body weight by either altering the internal reference value sought by the controller or by modulating the primary afferent signals regarding fat
30 stores that are analyzed by the controller (Bray G.A. et al., *Nature*. 404:672-674 (2000) and (Healtheon/WebMD. (1999)).

 While physical exercise and reductions in dietary intake of calories will dramatically improve the diabetic condition, compliance with this treatment is very poor because of well-entrenched sedentary lifestyles and excess food consumption, especially high fat-containing food.
35 Increasing the plasma level of insulin by administration of sulfonylureas (e.g. tolbutamide, glipizide) which stimulate the pancreatic β -cells to secrete more insulin or by injection of insulin after the response to sulfonylureas fails, will result in high enough insulin concentrations to stimulate the very

insulin-resistant tissues. However, dangerously low levels of plasma glucose can result from these last two treatments and increasing insulin resistance due to the even higher plasma insulin levels could theoretically occur. The biguanides increase insulin sensitivity resulting in some correction of hyperglycemia. However, the two biguanides, phenformin and metformin, can induce lactic acidosis and nausea/diarrhea, respectively.

Methods for Determining a Body Fat Value

Obesity is loosely defined as an excess of fat over that needed to maintain health, while it is formally defined as a significant increase above ideal weight, ideal weight being defined as that which maximizes life expectancy (Friedman, J.M. *Nature*. 404:633 (2000)). A convenient clinical and epidemiological measure of adiposity is the body mass index (BMI), which is calculated as weight divided by the square of the height (kg/m^2). BMI is highly correlated with more complex measures of body fat, such as those described herein, although the relation is less accurate at the extremes of the height distribution. (Healtheon/WebMD 1999).

Body Mass Index

In clinical practice, body fat is most commonly and simply estimated by using a formula that combines weight and height. The underlying assumption is most variation in weight for persons of the same height is due to fat mass, and the formula most frequently used in studies is body-mass index (BMI). A graded classification of obesity using BMI values provides valuable information about increasing body fatness. It allows meaningful comparisons of weight status within and between populations and the identification of individuals and groups at risk of morbidity and mortality. It also permits identification of priorities for intervention at an individual or community level and for evaluating the effectiveness of such interventions. However, BMI may not correspond to the same degree of fatness across different populations. Nor does it account for the wide variation in the nature of obesity between different individuals and populations (Kopelman P.G. *Nature*. 404:635 (2000)).

The World Health Organization provides the following classifications of overweight using BMI:

Table C

BMI (kg/m^2)	W.H.O. classification	Popular description
18.5	Underweight	Thin
18.5-24.9	normal	Healthy
25.0-29.9	Grade 1 overweight	Overweight
30.0-39.9	Grade 2 overweight	Obese
>40.0	Grade 3 overweight	Morbidly obese

Other Methods of Measuring a Body Fat Value

In addition to BMI, there are number of methods of determining fat mass measurements including waist circumference, waist-to-hip ratio, skinfold thickness, and bioimpedance (Heymsfield S.B. et al. *Am J Clin Nutr*. 64:478-84 (1996)) and (Calle E.C. et al. *New Engl J Med*. 341:1097-

1104 (1999) and (Gallagher D. et al. Am J Epidemiol. 143:228-39 (1996). Table D, herein, discusses each of these methods.

Table D

Method	Definition	Advantages/limitations
BMI	Weight in kilograms divided by square of the height in meters	BMI correlated strongly with densitometry measurements of fat mass: main limitation is that it does not distinguish fat mass from lean mass
Waist circumference	Measured (in centimeters) at midpoint between lower border of ribs and upper border of pelvis	Waist circumference measures for assessing upper body fat deposition: neither provide precise estimates of intra-abdominal (visceral) fat
waist-to-hip ratio	Ratio of the waist circumference and the hip circumference measured (in centimeters) at the upper border of pelvis	Waist-to-hip ratio is a good indicator of abdominal (i.e., android, as opposed to gynecoid) obesity, which is an even more important risk factor for NIDDM than obesity.
Skinfold thickness	Measurement of skinfold thickness (in centimeters) with callipers provides a more precise assessment if taken at multiple sites	Measurements are subject to considerable variation between observers, require accurate callipers and do not provide any information on abdominal and intramuscular fat
Bioimpedance	Based on the principle that lean mass conducts current better than fat mass because it is primarily an electrolyte solution: measurement of resistance to a weak current (impedance) applied across extremities provides an estimate of body fat using an empirically derived equation	Devices are simple and practical but neither measure fat nor predict biological outcomes more accurately than simpler anthropometric measurements

(Kopelman P.G. Nature. 404:635 (2000))

5 REFERENCES

1. Bundred, P., Kitchiner, D. & Buchan, I. Prevalence of overweight and obese children between 1989 and 1998: population based series of cross sectional studies. *Brit Med.J.* **322**, 313-314 (2001).
2. Taniguchi, A., Kono, T., Okuda, H., Oseko, F., Nagata, I., Kataoka, K., Imura, H. Neutral glyceride synthesis from glucose in human adipose tissue: growing and mature subjects. *J. Lip. Res.* **27**, 925-929 (1986).
3. Le Stunff, C., Fallin, D., Schork, N.J. & Bougnères, P. The insulin gene VNTR is associated with fasting insulin levels and development of juvenile obesity. *Nat Genet.* **26**, 444-446 (2000).
4. Kennedy, G.C., German, M.S. & Rutter, W.J. The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. *Nat Genet.* **9**, 293-298, (1995).
5. Paquette, J., Giannoukakis, N., Polychronakos, C., Vafiadis, P. & Deal, C. The INS 5' variable number of tandem repeats is associated with IGF2 expression in humans. *J. Biol. Chem.* **273**,

14158-64 (1998).

6. Reik, W. & Walter, J. Genomic imprinting : parental influence on the genome. *Nat. Rev.* **2**, 21-32 (2001).

7. Vafiadis, P. et al. Imprinted genotype-specific expression of genes at the IDDM2 locus in
5 pancreas and leucocytes. *J. Autoimmun.* **9**,397-403 (1996).

8. Bennett, S.T. et al. IDDM2-VNTR-encoded susceptibility to type 1 diabetes: dominant protection and parental transmission of alleles of the insulin gene-linked minisatellite locus. *J. Autoimmun.* **9**, 415-421 (1996).

9. Paquette, J., Giannoukakis, N., Polychronakos, C., Vafiadis, P. & Deal, C. The INS 5'
10 variable number of tandem repeats is associated with IGF2 expression in human. *J. Biol. Chem.* **273** :14158-14164 (1998).

10. Eaves, I.A. et al. Transmission ratio distortion at the INS-IGF2 VNTR. *Nat. Genet.* **22**,324-5 (1999).

11. Bennett, S.T. & Todd, J.A. Human type 1 diabetes and the insulin gene: principles of
15 mapping polygenes. *Annu. Rev. Genet.* **30**, 343-370 (1996).

12. Huxtable, S.J. et al. Analysis of parent-offspring trios provides evidence for linkage and association between the insulin gene and type 2 diabetes mediated exclusively through paternally transmitted class III variable number tandem repeat alleles. *Diabetes* **49**, 126-130 (2000).

13. Weinberg, C.R. Methods for detection of parent-of-origin effects in genetic studies of case-
20 parents triads. *Am J Hum Genet* **65**, 229-235 (1999).

14. Weinberg, C.R., Wilcox, A.J. & Lie, R.T. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am. J. Hum. Genet.* **62**, 969-978 (1998).

15. Schaid, D.J. Likelihoods and TDT for the case-parents design. *Genet. Epidemiol.* **16**, 250-260 (1999).

16. Sham, P.C. and D. Curtis, An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics*, **59**(Pt 3): p. 323-36. (1995).

17. Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C.G. & Polychronakos, C. Parental genomic imprinting of the human IGF2 gene. *Nat Genet.* **4**, 98-101 (1993).

18. Moore, G.E., et al., Evidence that insulin is imprinted in the human yolk sac. *Diabetes*, **50**(1): p. 199-203. (2001).

19. Lew, A., Rutter, W.J. & Kennedy, G.C. Unusual DNA structure of the diabetes susceptibility locus IDDM2 and its effect on transcription by the insulin promoter factor Pur-1/MAZ. *Proc. Natl. Acad.Sci. USA* **97**, 12508-12512 (2000).

20. Whitaker, R.C. & Dietz, W.H. Role of the prenatal environment in the development of obesity. *J. Pediatr.* **132**, 768-776 (1998).

21. Dunger, D.B., et al. Association of the INS VNTR with size at birth. *Nat. Genet.* **19**, 98-100 (1998).
22. Hattersley, A.T., Beards, F., Ballantyne, E., Appleton, M., Harvey, R. & Ellard, S. Mutations in the glucokinase gene of the fetus result in reduced birth weight. *Nat. Genet.* **19**, 268-270 (1998).
23. Catalano, P.M., Thomas, A.J., Huston, L.P. & Fung, C.M. Effect of maternal metabolism on fetal growth and body composition. *Diabetes Care* **21**, B85-B90 (1998)
24. Battaglia, F.C. & Thureen P.J. Nutrition of the fetus and the premature infant. *Diabetes Care* **21**, B70-B74 (1998)
25. Dietz, W.H. Critical periods in childhood for the development of obesity. *Am. J. Clin. Nutr.* **59**, 955-959 (1994).
26. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *History and Geography of Human Genes* (Princeton University Press, Princeton, (1994).
27. International Obesity Task Force. Obesity: Preventing and managing the global epidemic. Report of a WHO consultation on Obesity, 3-5 June 1998, Geneva: WHO(1998).
28. Freeman, J.V., Power, C., Rodgers, B. Weight-for-height indices of adiposity: Relationships with height in childhood and early adult life. *Int. J. Epidemiol.* **24**, 970-976 (1995).
29. Cole, T.J., Freeman, J.V., Preece, M.A. Body mass index reference curves for the UK. *Arch. Dis. Child* **73**, 25-29 (1995).

EXAMPLES

The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the present invention, and are not intended to limit the scope of what the inventors regard as their invention nor are they intended to represent that the experiments below are all or the only experiments performed. Efforts have been made to ensure accuracy with respect to numbers used (e.g. amounts, temperature, etc.) but some experimental errors and deviations should be accounted for. Unless indicated otherwise, parts are parts by weight, molecular weight is weight average molecular weight, temperature is in degrees Centigrade, and pressure is at or near atmospheric.

Example 1

Paternal transmission of the very common class I INS VNTR alleles predisposes Caucasian children to multifactorial obesity of early onset

METHODS

Subjects

The vast majority of the studied obese patients came from a previously described cohort (3) originating from Mediterranean and Central Europe countries. The geographic origin of the patients

was assessed through family history, analysis of patronymic names and grandparents birthplace (26). Mediterranean and Central Europeans had comparable multi-site insulin region haplotypes (determined from 6 neighbouring SNPs using haplotype estimation and likelihood ratio testing of equality between haplotype profiles), reflecting their close genetic origin (3). A subset of additional probands came from our ongoing recruitment since last report. From this cohort, we selected 402 Caucasian children whose onset of obesity occurred before 6 years of age, a critical period of childhood obesity development (27), and whose parents were available to sampling (Table 1).

We defined obesity onset arbitrarily (3) as the date at which, due to rapid and monotonic weight increase, the body mass index crosses the 85th percentile for age and sex. From these 402 obese children, 140 had a father who was heterozygous for the VNTR class I and III alleles, and 125 had a heterozygous mother (27 with both heterozygous parents, 238 total eligible probands, excluding non-informative trios containing heterozygous parent and child). These trios were all from different families. 121 lean siblings of these obese probands were also collected (Tables 1 & 2). We selected siblings older than 6 yrs of age to be absolutely sure that none of them had developed early obesity. Leanness was defined as a relative weight = 100% of a standard weight given height, age and sex (28).

Genotyping

The obese and lean children and their parents were genotyped at the VNTR locus as previously reported (3). -23 *HphI* SNP genotypes were determined by the analysis of PCR products. In Caucasians, *HphI* '+' alleles (A) are in complete LD with class I alleles of the neighboring VNTR, '-' alleles (T) with class III alleles (11) : only 0.23% haplotypes are discordant between *HphI* '+' and VNTR class I alleles (11). Therefore, we tested the insulin VNTR by using -23 *HphI* as a surrogate marker.

Genotyping was carried out as follows. Genomic DNA was subjected to PCR using the following primers: INS04: TCCAGGACAGGCTGCATCAG (SEQ ID NO:5); and INS05: AGCAATGGGCGGTTGGCTCA (SEQ ID NO:6). Typical PCR conditions: 96-well microtiter plates (Perkin), each 50µl reaction containing 200 ng DNA, 1.5 mM MgCl₂, 5 µl 10X reaction Buffer (Perkin Elmer), 10% DMSO (Pst1), 0.2 mM each dNTP, 1 µM of each primer and 1.25 U of Taq Polymerase (Perkin Elmer). 30-35 cycles were performed using a 9700 Perkin Elmer thermocycler. Using the INS04 and INS05 primers, with an annealing temperature of 65°C, a 441 bp PCR product is obtained. 10 µl of PCR products were digested with 2.5 U of *HphI* and gel electrophoresed to determine genotype. The [+] alleles indicate the restriction enzyme cuts the sequence, whereas [-] alleles indicate a cut was not made. ++ individuals give bands of 232, 161, and 39 bp; +/- individuals give bands of 271, 232, 161, and 39 bp; -/- individuals give bands of 271 and 161 bp.

TDI and parent-of-origin effects.

Transmission disequilibrium was first assessed via simple tabulation of transmitted and non-transmitted alleles and comparison of discordant transmissions (19). The estimated probability of transmission of the class I allele in particular (π) was compared between heterozygous mother-child pairs and heterozygous father-child pairs via chi-squared tests. Conditional logistic regression of VNTR allele transmission, matched on child-parent pairs, was expressed as:

$$\ln\left(\frac{P(t)}{1-P(t)}\right) = \alpha A + \beta A * F \text{ [where } P(t) = \text{probability that an allele was transmitted to the affected}$$

child; $A = 1$ if the allele is a class I allele and 0 otherwise ; $F = 1$ if the transmitted allele is derived from the father]. The only pairs relevant in this analysis are those involving heterozygous parents (discordant transmissions). Likelihood ratio testing was then used to assess significance of parent-of-origin effects by comparing the full model versus one with β constrained to 0.

For the parent-of-origin likelihood ratio test of Weinberg (PO-LRT), the data were categorized into 5 trio classes that are informative for parent-of origin and maternal genotype effects (see Table 3 of Weinberg, 1999). Within a particular trio category, an estimate of parent-of-origin effects can be observed as a positive log odds of paternal vs maternal origin of transmitted allele. This conditional scenario can be set up using indicator variables such that each of these 5 types of families contribute uniquely to the following (unconditional) logistic regression:

$$\ln\left(\frac{P(F > M \mid triotype)}{1 - P(F > M \mid triotype)}\right) = \alpha C + \beta_{s2} I_{P>1} + \beta_{s1} [I_{P=1} - I_{P>2}]$$

where M = number of class 1 alleles in mother's genotype; F = number of class 1 alleles in father's genotype; $C = 1$ if child is heterozygous, 0 otherwise; $P = M+F$; I = parental class I allele sum; I = indicator if subscript statement is true, 1 = Y, 0 = N. The coefficients of this regression can be interpreted as $\alpha = \ln(I_F)$ where I_F = increased risk for paternally-derived class I alleles versus maternally-derived ones; $\beta_{s1} = \ln(S_1)$ where S_1 is the relative risk for a child whose mother has one copy of the class I allele versus those with III/III mothers; and $\beta_{s2} = \ln(S_2)$ where S_2 is the relative risk for having a mother with I/I genotype (versus III/III). Tests for parent-of-origin (PO) effects or maternal genotype effects can be carried out via nested models using likelihood ratio tests.

The alternative log linear approach expresses the expected number of trios in one of the 15 possible joint mother, father, child genotype categories as a log-linear function of the genotype risk (in proband), the maternal genotype risk, and parent-of-origin effects:

$$\ln[E(n_{MPC})] = \beta_1 C_1 + \beta_2 C_2 + \gamma_1 M_1 + \gamma_2 M_2 + \alpha F + \ln(\eta) I_{MPC}$$

where n_{MPC} reflects the number of trios in a particular 'MPC' category where M,P,C represents the number of class I alleles carried by the mother, father, and child respectively, $C_1 = 1$ if $C=1$; $C_2 = 1$ if $C=2$; $M_1 = 1$ if $M=1$; $M_2 = 1$ if $M=2$; $F = 1$ if $C=1$ and was derived from the father; $I_{MPC} = 1$ if $M=P=C=1$. The coefficients can be interpreted as $\alpha = \ln(I_F)$ where I_F = increased risk for paternally-derived class I alleles versus

maternally-derived ones; $\beta_1 = \ln(R_1)$ where R_1 is the relative risk for a child with 1 copy of allele I versus III/III children; $\beta_2 = \ln(R_2)$ where R_2 is the relative risk for a child with I/I genotype compared to III/III types; $\gamma_1 = \ln(S_1)$ where S_1 is the relative risk for a child whose mother has one copy of the class I allele versus those with III/III mothers; and $\gamma_2 = \ln(S_2)$ where S_2 is the relative risk for having a mother with I/I genotype versus III/III mothers. Likelihood ratio tests of nested models were used to test parent-of-origin and maternal genotype effects.

All analyses were run on SAS version 10.0. Conditional logistic analysis was accomplished in proc logistic using the no intercept option and adjusting outcome and dependent variables accordingly. The log-linear models were performed in proc genmod.

10 RESULTS

We have recently reported finding no direct association between these VNTR alleles and childhood obesity, based on the observation of no difference in insulin VNTR genotype distributions between these obese children and lean controls (3). In the present report, we have investigated the possibility of parent-of-origin differences in transmission of the VNTR allele classes to obese children, using a case-parent trio design. Our previous case-control approach could not detect a class I allele effect because of the dilution of paternal class I effect by noncontributory maternal meioses.

To now be able to distinguish association with paternal and/or maternal alleles, we genotyped nuclear families consisting of (young) obese offspring, their lean sibs whenever possible (Table 1), and both parents.

Table 1. Characteristics of the obese children and their lean siblings (Mean $\pm$ SD)

	Obese children	Lean siblings
N	238	106
Sex (M/F)	98/140	66/40
Age at study (yrs)	11.5 $\pm$ 3.1	15.2 $\pm$ 5.1
BMI (kg/m ²)	29.2 $\pm$ 3.5	18.1 $\pm$ 2.1
centile*	> 99 th	26.5 $\pm$ 4.1
Age of obesity onset (yrs)**	4.7 $\pm$ 1.5	-
* adjusted to age and sex (29)		
** onset defined as BMI $\geq$ 85th ctile for age (see Methods. Subjects)		

Table 2 shows the distribution of heterozygous mothers and fathers, the number of transmitted class I alleles to obese children, and the estimated proportion of transmission of class I allele (π) for the overall and parent-of-origin subsets. To our surprise, we found a large excess of paternal transmission of class I versus class III alleles (Table 2). The estimated relative risk of early-onset obesity for children inheriting a class I allele from their father is $r_{If} = 1.8$. No transmission distortion was seen, however, from either parent type to lean siblings, in the same data set (see Table 2, right).

Table 2. Number of heterozygous parents by gender and number of transmitted insulin VNTR class 1 alleles to obese children and to lean siblings.								
	Obese children				Lean siblings			
	# Htz parents*	# T**	π	π^2 (P val)	# Htz parents*	# T**	π	π^2 (P val)
TDT_{MvF}								
Mothers	125	57	0.456		61	30	0.492	
Fathers	140	90	0.643		60	33	0.550	
Overall	265	147	0.555	9.3 (.002)	121	63	0.521	.41 (.52)
TAT								
Mothers	93	38	0.409		46	21	0.467	
Fathers	108	71	0.657		45	24	0.543	
Overall	201	109	0.540	12.4 (.0004)	91	45	0.495	0.53 (.47)

* Number of heterozygous parents does not include parents from trios where all three members are heterozygous, as these are uninformative. ** T = transmitted insulin VNTR class I allele; π =estimated probability of transmission of a class I allele to the child; π^2 = test of $\pi_M = \pi_F$.

5

In a recent paper, Weinberg et al proposed that the test of homogeneity between π_m and π_f (Table 2 top, TDT_{MvF}) may be biased because trios of doubly heterozygous parents contribute twice, while all singly heterozygous parent trios only contribute once to the π calculations (13,14). As an alternative, she proposed using only trios with a single heterozygous parent in the analysis (transmission asymmetry test, TAT). Our data show evidence for significant excess paternal transmission using this method as well (Table 2, bottom).

10

The TDT scenario can be also be expressed in a likelihood framework and likelihood ratio testing can be used to test for differential effects of transmission of risk alleles from fathers versus mothers (13-15). We present three approaches to likelihood-based parent-of-origin tests Table 3.

15

First, the TDT can be framed as a conditional logistic regression (grouped by parent-child pair) (16) with models including or excluding allele parent-of-origin as a covariate. The likelihood ratio test between these models for the obese child trios (Table 3, A) shows a significant effect of the inclusion of the parent-of-origin term. This test was not significant among the lean sibling-parent trios.

Table 3. Tests for parent-of-origin (PO) and (non-transmitted) maternal genotype effects for obese child-parent trios and lean sibling-parent trios.										
A. CONDITIONAL LOGISTIC REGRESSION * $\ln[P(Transmission)/(1-P(Trans))]=\alpha A+\beta A * F$ (conditional on pair)										
Obese probands :										
Model	Estimates		-2LL	Test	LRT χ^2	df	P val			
	a	β								
Allele effect (A)	1.250	-	364.19	PO	9.38	1	0.002			
Allele and PO** (A + A*F)	-0.177	0.764	354.81							
Lean sibs:										
Allele effect (A)	1.086	-	167.54	PO	0.42	1	0.521			
Allele and PO (A + A*F)	0.201	-0.234	167.12							
B. PO-LRT METHOD*										
$\ln\left(\frac{P(F > M \text{trio type})}{1 - P(F > M \text{trio type})}\right) = \alpha C + \beta_{s1} I_{F>1} + \beta_{s2} [I_{F=1} - I_{F>2}]$										
Obese probands :										
Model	Estimates			-2LL	Test	LRT χ^2	df	P val		
	χ	β_{s1}	β_{s2}							
Full (PO, maternal genotypes)	2.72	0.82	0.47	264.52		-	-	-		
Maternal genotypes only	-	1.56	1.45	276.43	PO	11.91	1			
PO only	1.69	-	-	271.62	Maternal genotype effects	7.10	2			
C. LOG-LINEAR METHOD*										
$\ln[E(n_{MPC})] = \beta_1 C_1 + \beta_2 C_2 + \gamma_1 M_1 + \gamma_2 M_2 + \alpha F + \ln(\eta) I_{MPC}$										
Obese probands :										
Model	β_1	β_2	γ_1	γ_2	a	-2LL	Test	LRT χ^2	df	P val
Full (child and maternal genotypes, PO)	0.20	0.13	0.77	1.20	1.05	872.12	PO	21.88	1	<0.001
Child and maternal genotypes only	0.77	1.28	0.19	0.59	-	850.24				
Child genotypes and PO only	0.60	1.06	-	-	0.66	855.09	Maternal genotype effects	17.03	2	<0.001

* See methods for interpretation of coefficients as risk parameters; **PO=parent of origin effects.

- 5 Two recent papers by Weinberg (13) and Weinberg et al (14) pointed out that non-transmitted maternal genotype effects, such as *in utero* effects, could significantly modify the distribution of maternal alleles among affected probands, leading to false interpretations regarding parent-of-origin effects on transmission as presented in Table 2 and Table 3, A. As an alternative, she proposed two approaches that model maternal genotype effects separately from transmission

effects and parent-of-origin effects. With this in mind, our second likelihood-based approach is a conditional logistic regression method which models the probability that the father's class I allele was transmitted to an affected child (versus the mother's) as the outcome, conditioned on mating type and child's genotype (13). This approach also shows a significant parent-of-origin effect (excess transmission of paternal alleles), as well as evidence for (non-transmitted) maternal genotype effects (Table 3, B).

Our third likelihood-based approach considers the distribution of trio types (defined by the genotypes of the mother, father, and child) in the data set as a 15-nomial. The expected number of trios per category is expressed as a log-linear function of the number of risk alleles transmitted to the child, the number carried by the mother, and the parent-of-origin of transmitted alleles. Likelihood ratio tests under this framework also showed evidence for paternal transmission as well as (non-transmitted) maternal genotype effects. In summary, each method showed evidence for a paternal transmission effect of the class I allele. This effect was not observed using a similar set of models for the lean siblings, further implicating excess paternal transmission of the class I insulin alleles in risk for childhood obesity (Table 3). Further, the proband gender distribution was not significantly different between transmitted and non-transmitted groups ($p=0.12$), nor was there evidence for a difference in paternal age between transmission groups among the heterozygous father-child pairs ($p=0.66$), suggesting proband gender and paternal age do not influence these results.

We think that our observations relate to the regulation of the *in utero* expression of insulin and IGF2 genes, two major regulators of fetal growth that are known to be maternally imprinted (17,18). We have shown previously that class I VNTR alleles are associated with increased insulin secretion in obese children, at an age when insulin gene expression is bi-allelic (3). It was known from *in vitro* studies that class I alleles are associated with increased transcription of the insulin gene (4), possibly through formation of specific G-rich DNA structures (19). In the fetal pancreas, class I VNTR alleles are associated with increased transcription of the insulin gene (7,8). The present results suggest a role of the imprinted VNTR-INS-IGF2 region in the predisposition to early obesity. Both insulin and IGF2 may promote adipogenesis and/or lipid storage at the end of gestation in humans (20). Pre- and post-natal growth is affected by the genetic modulation of insulin secretion (21,22).

In addition to paternal effects, our analysis shows the existence of a smaller maternal VNTR genotype effect. We found that maternal class III VNTR alleles are associated with an increased risk of obesity in the offspring. This effect, which is not related to maternal allele transmission, is independent of the child's genotype (Table 3). There likely are interactions in the childbearing mothers between their VNTR genotype, control of insulin secretion and the metabolic constraints of pregnancy, characterized by a high degree of insulin resistance. Previous studies have reported lower insulin secretion (3) and increased risk of non insulin dependent diabetes (11) in women with class III VNTR alleles. In addition, about 32% of mothers in our sample were obese before starting pregnancy. Depending on the maternal VNTR, there could be changes in the maternal milieu and

materno-fetal glucose homeostasis (23,24). This hypothesis will be tested in future studies. There is support from the literature that the third trimester of pregnancy may represent a critical period for the entrainment of postnatal fatness (25).

5 In conclusion, our observations suggest that a programming of human fetuses, primarily through mechanisms related to paternal VNTR alleles and expression of neighbouring gene(s), could be a widespread mechanism predisposing to common obesity of early onset. Other genetic and non genetic factors, including maternal metabolic and nutritional status, likely interfere with these mechanisms.

10 While the present invention has been described with reference to the specific embodiments thereof, it should be understood by those skilled in the art that various changes may be made and equivalents may be substituted without departing from the true spirit and scope of the invention. In addition, many modifications may be made to adapt a particular situation, material, composition of matter, process, process step or steps, to the objective, spirit and scope of the present invention. All
15 such modifications are intended to be within the scope of the claims appended hereto.

CLAIMS

What is claimed is:

1. A method of determining the risk of developing obesity in an individual, comprising
5 determining a paternal insulin variable number of tandem repeats (VNTR) allele in the individual by
determining the identity of a polymorphic base of at least one marker in linkage disequilibrium with
the insulin VNTR of the individual, wherein the presence of a paternal insulin VNTR class I allele
indicates that the individual has an approximately two-fold increase in risk of developing obesity
compared to an individual carrying a paternal insulin VNTR class III allele.
10
2. A method of treating obesity in an individual, comprising administering a weight
loss or a weight control regimen in an individual identified by a method according to claim 1 as being
at risk of developing obesity, thereby treating obesity in the individual.
- 15 3. A method of reducing the risk that an individual will develop an obesity-related
disorder, comprising administering a weight loss or a weight control regimen in an individual
identified by a method according to claim 1 as being at risk of developing obesity, thereby reducing
the risk that the individual will develop an obesity-related disorder.
- 20 4. The method of any one of claims 1-3, wherein the marker is -23 HphI.

SEQUENCE LISTING

<110> Bougneres, Pierre

<120> Methods for assessing the risk of
obesity based on allelic variations in the 5'-flanking
region of the insulin gene

<130> BOUG-002WO

<140> Unassigned

<141>

<150> 60/309,235

<151> 2001-07-31

<150> 60/316,830

<151> 2001-08-31

<160> 30

<170> FastSEQ for Windows Version 4.0

<210> 1

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 1

tgacgccaaag gacaagctca

20

<210> 2

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 2

ccagcagccc cagtcctgca

20

<210> 3

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 3

caccagctgg ccttcaaggt

20

<210> 4

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 4

gctgggcact aacaaggtgt

20

<210> 5

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 5

tccaggacag gctgcatcag

20

<210> 6

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 6

agcaatgggc ggttggctca

20

<210> 7

<211> 18

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 7

taaagccctt gaaccagc

18

<210> 8

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 8

cagcccagcc tcctccctcc aca

23

<210> 9

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic primer

<400> 9

cccaggggcc gaagagtca

19

<210> 10

<211> 20

<212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 10
 gctgagctgg cagcgattca 20

 <210> 11
 <211> 22
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 11
 cttggacttt gagtcaaatt gg 22

 <210> 12
 <211> 20
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 12
 cctcctttgg tcttactggg 20

 <210> 13
 <211> 13
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 13
 acaggggtgt ggg 13

 <210> 14
 <211> 14
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 14
 gtggggacag gggt 14

 <210> 15
 <211> 15
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 15

cctggggaca ggggt 15

<210> 16
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 16
ctggggacag ggggt 14

<210> 17
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 17
ccggggacag ggggt 14

<210> 18
<211> 15
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 18
cccggggaca ggggt 15

<210> 19
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 19
gtggggatag ggggt 14

<210> 20
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 20
gtgggcacag ggggt 14

<210> 21
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
 <223> synthetic primer

 <400> 21
 ctgctgagga cttgctgctt g 21

 <210> 22
 <211> 21
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 22
 cagaaggaca gtgatctggg t 21

 <210> 23
 <211> 34
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 23
 tcatgcggtcc atggtccgga acccctgtcc ccac 34

 <210> 24
 <211> 35
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 24
 tcatgcggtcc atggtccgga acccctgtcc ccagg 35

 <210> 25
 <211> 34
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 25
 tcatgcggtcc atggtccgga acccctgtcc ccag 34

 <210> 26
 <211> 34
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> synthetic primer

 <400> 26
 tcatgcggtcc atggtccgga acccctgtcc ccgg 34

 <210> 27

<211> 34
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 27
tcatgcggtcc atggtccgga acccctatcc ccac

34

<210> 28
<211> 35
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 28
tcatgcggtcc atggtccgga acccctgtcc ccggg

35

<210> 29
<211> 34
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 29
tcatgcggtcc atggtccgga acccctgtgc ccac

34

<210> 30
<211> 36
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic primer

<400> 30
ctgctgagga cttgctgctt gcaggggtgt ggggat
1

36