

METHODS FOR ASSESSING WHETHER A GENETIC REGION IS ASSOCIATED WITH INFERTILITY

Related Application

5 This application claims the benefit of and priority to U.S. Provisional No. 61/932,233,
filed January 27, 2014, which is incorporated by reference in its entirety.

Technical Field

10 The invention generally relates to methods for assessing whether a genetic region is
associated with fecundity and fertility disorders.

Background

15 Approximately one in seven couples has difficulty conceiving. Infertility may be due to a
single cause in either partner, or a combination of factors (e.g., genetic factors, diseases, or
environmental factors) that may prevent a pregnancy from occurring or continuing. Every
woman will become infertile in her lifetime due to menopause. On average, egg quality and
number begins to decline precipitously at 35. However, some women experience this decline
much earlier in life, while a number of women are fertile well into their 40s. Similarly, while it is
normal for women's reproductive lifespans to include periods of natural infertility, associated
20 with menstrual periods or post-partum changes in reproductive endocrinology, for example,
some women experience abnormally extended periods of infertility. Such disorders are referred
to as infertility-, fecundity-, or fertility-related disorders. Though, generally, advanced maternal
age (35 and above) is associated with poorer fertility outcomes, there is no way of diagnosing
egg quality issues in younger women or knowing when a particular woman will start to
25 experience decline in her egg quality or reserve.

 The elucidation of the genetic basis of female fecundity and fertility disorders permits the
development of powerful, rapid, and non-invasive diagnostic tools that will help clinicians direct
patients to efficient and effective treatment options. Additionally, the discovery of the key
genetic loci underlying these disorders holds great promise for the identification of novel targets
30 for drug development and therapeutics. Finally, a better understanding of the crucial molecular

pathways underlying human fecundity and fertility guides the next generation of targeted, non-hormonal contraceptives.

Summary

5 The invention utilizes the status of various fecundity and fertility-related genomic regions in order to assess risk and/or susceptibility to reduced fecundity, fertility, premature menopause, or extended periods of infertility. Methods of the invention utilize genomic information, including, but not limited to, one or more polymorphisms in one or more fecundity- or fertility-related genomic regions, mutations in one or more of those regions, or epigenetic factors
10 affecting expression in those regions. Mutations in a fecundity- or fertility-related genomic region may result in an alternative splicing event, lowered or increased RNA expression, and/or alterations in protein expression, with concomitant physiological changes. Methods of the invention are useful for informing a patient of her susceptibility to abnormally extended periods of infertility or reduced fecundity in connection with age or other relevant phenotypic factors,
15 such as hormone levels or ovarian follicle count.

 The invention generally provides methods for assessing whether a genomic region is associated with a fertility-related condition. Aspects of the invention are accomplished using a transgenic animal, such as a genetically-modified mouse. A genomic region suspected to be associated with abnormal fecundity or extended period of infertility is identified. Using that
20 information, the invention provides for genomic modification of a test animal, such as a mouse. The genetically-modified animal is then assessed for the presence of an infertility-associated phenotype. The presence of the phenotype is indicative that the selected genomic region is associated with an infertility-related condition. Methods of the invention allow for the discovery of the key genomic regions underlying fecundity, fertility and infertility and for the subsequent
25 identification of novel targets for drug development and therapeutics. Additionally, genetically-altered test animals that show presence of an infertility phenotype are useful for therapeutic testing.

 A genetic locus can encompass a gene and/or upstream and downstream elements, such as introns, promoters and the like, that are involved in the expression of that gene or other
30 genetic loci. There are numerous methods that are useful to identify a genetic locus whose function is suspected of being associated with extended infertility, including reference to

literature, databases and empirical analysis. In certain embodiments of the invention, identifying a fertility-related genomic region involves obtaining data on a set of genetic loci, the set including loci known to be associated with infertility and loci having no prior association with infertility. A clustering analysis is then performed on the data to identify genetic loci that have
5 no prior association with infertility that cluster with one or more genetic loci known to be associated with infertility. Thus, genetic loci that have no prior association with infertility are identified as being infertility-related by virtue of clustering with known infertility-related genetic loci. For example, a genetically-altered mouse having a gene knock-out is produced to determine if that gene is implicated in an infertility-associated phenotype. In that manner,
10 genetic loci not previously associated with infertility are identified as potential infertility biomarkers.

Infertility may not be the result of a single genomic alteration, but rather may be the result of a combination of multiple factors or multiple alterations. Methods of the invention provide a better understanding of the molecular pathways underlying human fertility. For
15 example, presence of an infertility-associated phenotype is used as a factor in ranking the importance of a gene in a database of genetic loci associated with infertility in humans by associated the gene (or more often a mutation) with the phenotype. A correlation between the presence of an allele or a mutation in a gene with phenotype increases or decreases the predictive value of the contribution of the genomic region to phenotype.

20 Additionally, the invention provides genetically altered mice for testing therapeutic agents. In those embodiments, methods of the invention further involve administering a therapeutic agent to the mouse, and assessing the effect of the therapeutic agent on phenotype. A therapeutic agent that rescues the phenotype, i.e., returns or partially re-establishes the wild type fertility phenotype, is a good drug candidate.

25 Other aspects of the invention provide methods for assessing whether a human genomic alteration is associated with an infertility phenotype in a mouse. Those methods involve identifying a human genomic region whose function is known to be associated with human infertility. The methods additionally involve producing a genetically-modified mouse in which the genetic region whose function is associated with human infertility is altered. The mouse is
30 then assessed for presence of the infertility phenotype.

Other aspects and alternatives for use of the present invention are apparent to the skilled artisan as provided in the detailed description of the invention that follows.

Brief Description of the Drawings

5 Fig. 1 depicts the rate of decline of fertility with age and the corresponding increase in the risk of infertility with age. The shaded areas represent different age groups who would benefit from a genetic screen for infertility risk (late teen to mid 40's) versus a genetic screen of premature decline in fertility (late teens to late 30's).

Fig. 2 depicts one way that phenotypic variables can be utilized to accelerate the
10 discovery of genetic regions related to female infertility.

Fig. 3 depicts the methodology for integrating clinical data with genomic data to predict treatment dependent and independent fertility outcomes.

Fig. 4 depicts the different kinds of genetic variants associated with risk of infertility.

Fig. 5 depicts a method for filtering through variants detected in whole genome
15 sequencing for the identification of genetic regions related to infertility.

Fig. 6 depicts some of the components of the FertilomeTMDatabase, a tool for correlating genetic regions with risk for infertility (FertilomeTMScore).

Fig. 7 is the bioinformatics pipeline used to identify biologically interesting and statistically significant genetic variants in infertile patients.

20 Fig. 8 shows the different types of biologically or statistically significant genetic variants that were detected in infertile patients in the MUC4 genetic region.

Fig. 9 provides CGH array data of copy number variations associated with infertility.

Fig. 10 illustrates a specific copy number variation detected in the GJC2 gene of
Chromosome 1.

25 Fig. 11 illustrates a specific copy number variation detected in the CRTC1 and GDF1 genes of Chromosome 19.

Fig. 12 illustrates a specific copy number variation detected in a non-coding region of Chromosome 6.

Fig. 13 illustrates population stratification correction of two patient groups (ZA = patients
30 who did not get pregnant with IVF treatment, ZB= patients with infertility who did get pregnant with IVF treatment).

Fig. 14 depicts an area of the cluster analysis results.

Fig. 15 illustrates a system for implementing methods of the invention.

Detailed Description

5 The invention generally relates to methods for the identification and determination of genetic loci and phenotypic characteristics related to infertility in humans and mice to develop a mouse model. Furthermore, the information gained from the present invention may be used in generating a mouse model for therapeutic investigations in infertility in humans. The invention generally relates to data analysis of genetic loci and phenotypes to determine not only the
10 relationship between genetic loci and phenotypic characteristics in a mammalian species, but also to identify genetic loci and corresponding phenotypes that are expressed in both humans and mice. By employing ranking methodologies, biomarkers, or genetic loci, that are expressed in both humans and mice can be determined. The present invention provides a powerful data set to be used in development of a mouse model for therapeutic investigations and strategy
15 development in human infertility.

Biomarkers

 A biomarker generally refers to a molecule that may act as an indicator of a biological state. Biomarkers for use with methods of the invention may be any marker that is associated
20 with infertility. Exemplary biomarkers include genes (e.g., any region of DNA encoding a functional product), genetic regions (e.g., regions including genes and intergenic regions with a particular focus on regions conserved throughout evolution in placental mammals), and gene products (e.g., RNA and protein). In certain embodiments, the biomarker is an infertility-associated genetic region. An infertility-associated genetic region is any DNA sequence in which
25 variation is associated with a change in fertility. Examples of changes in fertility include, but are not limited to, the following: a homozygous mutation of an infertility-associated genetic locus leads to a complete loss of fertility; a homozygous mutation of an infertility-associated genetic locus is incompletely penetrant and leads to reduction in fertility that varies from individual to individual; a heterozygous mutation is completely recessive, having no effect on fertility; and the
30 infertility-associated genetic locus is X-linked, such that a potential defect in fertility depends on whether a non-functional allele of the genetic locus is located on an inactive X chromosome

(Barr body) or on an expressed X chromosome.

According to certain aspects, methods of the invention provide for determining infertility genetic regions of interest based on data obtained from public and private fertility/infertility related databases. Infertility/fertility related data may include genetic loci involved in the regulation of implantation, idiopathic infertility genetic loci, polycystic ovary syndrome (PCOS) genetic loci, egg quality genetic loci, endometriosis genetic loci, and premature ovarian failure genetic loci. As described below, the infertility/fertility related data can then be processed using evolutionary conservation to identify genomic regions and variations of interest.

Evolutionary conservation analysis involves, generally, comparing nucleic acid sequences among evolutionary and distantly related genomes to identify similarities and differences between coding and/or non-coding regions across the genomes. The similarity between a region being examined and the related genomes correlates to a degree of conservation. Regions (e.g., coding, non-coding regions, and intergenic regions flanking a gene) that maintain a high degree of similarity across genomes over time are considered highly conserved.

Differences between the examined region and regions of related genomes indicate that the examined region has evolved over time. If the examined region is conserved among related genomes, the region is generally considered to exhibit or perform functions that are important for the species (i.e., functionally relevant). This is because genetic abnormalities at functionally important regions are typically harmful to the species, and are phased out over the evolutionary time span. Because functional elements are subject to selection, functional regions tend to evolve at slower rates than nonfunctional regions. A degree of conservation (e.g., degree of similarity between a target genomic region and related genomes) that is considered to be functionally relevant depends on the particular application. For example, a functionally relevant degree of conservation may be 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, etc. Regions of genetic loci identified by evolutionary conservation as being functionally relevant can then be used as regions of interest for diagnosing diseases and disorders, such as infertility.

According to certain embodiments, infertility regions of interest are identified by performing evolutionary conservation analysis of one or more genetic loci obtained from infertility and/or fertility-related data. The process of filtering through infertility/fertility related databases using evolutionary conservation, according to the invention, is called the ABCoRE

algorithm. For example, nucleic acid data obtained from the infertility/fertility related databases can be compared to distantly related genomes in order to assess conservation of the infertility-related nucleic acid. Regions of the nucleic acid determined to be conserved are classified as infertility regions of interest. In one embodiment, methods of the invention assess conservation of coding regions to determine infertility regions of interest. In another embodiment, methods of the invention assess conservation of non-coding regions to determine infertility regions of interest. In further embodiments, methods of the invention assess conservation of intergenic regions (i.e., a non-coding region flanking a gene) to determine infertility regions of interest. In other embodiments, conservation of both coding and non-coding regions is assessed to determine infertility regions of interest. In any of the above embodiments, coding, non-coding, and intergenic regions may be classified as an infertility region of interest if they have a degree of conservation of, for example, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, etc.

In particular aspects, the following method is employed to determine whether a genomic region is a fertility region of interest using conservation analysis. First, private and/or public nucleic acid data corresponding to infertility or fertility is obtained. Next, one or more genetic loci from that data is examined for conservation. The coding regions (i.e., exons) of a gene, non-coding regions of the gene, and/or regions flanking the gene (intergenic regions upstream and downstream from the gene being examined) are then analyzed for conservation. According to certain embodiments, if the coding region is found to be conserved (e.g., a degree of conservation 90% or above), the coding region is considered to be an infertility region of interest. The degree of conservation of the non-coding region is then compared to the degree of conservation of the coding region. If the degree of conservation of the non-coding region is similar to the degree of conservation of the coding region, then the non-coding region is also classified an infertility region of interest. This degree of conservation comparison may also be used to determine whether intergenic regions flanking a gene should be classified as an infertility region of interest.

Conservation of coding and/or non-coding sequences is described in Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* 7: 959-966; Brenner, S., Venkatesh, B., Yap, W.H., Chou, C.F., Tay, A., Ponniah, S., Wang, Y., and Tan,

Y.H. 2002. Conserved regulation of the lymphocyte-specific expression of *lck* in the Fugu and mammals. *Proc. Natl. Acad. Sci.* 99: 2936–2941; Karolchik, Donna, et al. "Comparative genomic analysis using the UCSC genome browser." *Comparative Genomics*. Humana Press, 2008. 17–33; Santini, Simona, Jeffrey L. Boore, and Axel Meyer. "Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters." *Genome research* 13.6a (2003): 1111–1122; Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16: 939–945; and Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12: 739–748.

In particular embodiments, the infertility-associated genetic region is a maternal effect gene. Maternal effects genes are genetic loci that have been found to encode key structures and functions in mammalian oocytes (Yurttas et al., *Reproduction* 139:809–823, 2010). Maternal effect genes are described, for example in, Christians et al. (*Mol Cell Biol* 17:778–88, 1997); Christians et al., *Nature* 407:693–694, 2000); Xiao et al. (*EMBO J* 18:5943–5952, 1999); Tong et al. (*Endocrinology* 145:1427–1434, 2004); Tong et al. (*Nat Genet* 26:267–268, 2000); Tong et al. (*Endocrinology*, 140:3720–3726, 1999); Tong et al. (*Hum Reprod* 17:903–911, 2002); Ohsugi et al. (*Development* 135:259–269, 2008); Borowczyk et al. (*Proc Natl Acad Sci U S A.*, 2009); and Wu (*Hum Reprod* 24:415–424, 2009). The content of each of these is incorporated by reference herein in its entirety.

The above-described infertility genetic regions of interest may then be ranked according to significance using one or more the following ranking schemes of the invention.

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 1 below. In Table 1, HGNC (<http://www.genenames.org/>) reference numbers are provided when available.

Table 1 below depicts one possible gene ranking scheme for the relative infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. The number of variants column corresponds to the experimental observations of these variants in a study of women with unexplained infertility. The most highly ranked (from top to bottom) genes in this list contained the most variants that were predicted to significantly affect protein structure and function (biologically significant) out of a list of fertility

related genes. Genetic variants considered to be biologically significant include mutations that result in a change: 1) to a different amino acid predicted to alter the folding and/or structure of the encoded protein, 2) to a different amino acid occurring at a site with high evolutionarily conservation in mammals, 3) that introduces a premature stop termination signal, 4) that causes a stop termination signal to be lost, 5) that introduces a new start codon, 6) that causes a start codon to be lost, 7) that disrupts a splicing signal, 8) that alters the reading frame or 9) that alters the dosage of encoded protein or RNA. All genetic variants detected from re-sequencing exclude sites where the variant allele is detected in only one chromosome (singletons) and sites sequenced in only one individual.

Table 1: Genomic loci containing biologically significant mutations ranked based on number of biologically significant variants observed in a study of unexplained female infertility.

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	Number of Variants detected	Variant Description (type and count)
MUC4	CMX-G0000006719	4585	7514	353	Drastic nonsynonymous:352; Start codon gained:1
EPHA8	CMX-G0000000415	2046	3391	23	CNV loss:23
LOXL4	CMX-G0000016263	84171	17171	11	CNV loss:11
FGF8	CMX-G0000016316	2253	3686	4	CNV gain:4
KISS1R	CMX-G0000026560	84634	4510	4	CNV gain:4
SCARB1	CMX-G0000019991	949	1664	4	Drastic nonsynonymous:1; Start codon gained:3
BARD1	CMX-G0000004834	580	952	3	Drastic nonsynonymous:1; Start codon gained:1; Start codon lost:1
DDX20	CMX-G0000001412	11218	2743	3	Start codon gained:3
ECHS1	CMX-	1892	3151	3	CNV gain:2; CNV

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	Number of Variants detected	Variant Description (type and count)
	G0000016594				loss:1
FMN2	CMX-G0000002910	56776	14074	3	Start codon gained:3
FOXO3	CMX-G0000010672	2309	3821	3	CNV gain:3
HS6ST1	CMX-G0000004221	9394	5201	3	Drastic nonsynonymous:3
MAP3K2	CMX-G0000004205	10746	6854	3	CNV gain:3
MST1	CMX-G0000005619	4485	7380	3	Drastic nonsynonymous:2; Splice site acceptor:1
MTRR	CMX-G0000008130	4552	7473	3	Drastic nonsynonymous:3
NLRP11	CMX-G0000028188	204801	22945	3	Drastic nonsynonymous:2; Start codon gained:1
NLRP14	CMX-G0000016919	338323	22939	3	Drastic nonsynonymous:3
NLRP8	CMX-G0000028191	126205	22940	3	Drastic nonsynonymous:2; Stop codon lost:1
ASCL2	CMX-G0000016707	430	739	2	Start codon gained:1; CNV gain:1
BMP6	CMX-G0000009564	654	1073	2	CNV loss:2
BRCA1	CMX-G0000025305	672	1100	2	Drastic nonsynonymous:2
BRCA2	CMX-G0000020222	675	1101	2	Drastic nonsynonymous:2
CENPI	CMX-G0000031175	2491	3968	2	Start codon gained:2
COMT	CMX-G0000029621	1312	2228	2	Drastic nonsynonymous:1; Start codon gained:1
CYP11B1	CMX-G0000013888	1584	2591	2	CNV gain:2
DAZL	CMX-G0000005296	1618	2685	2	Start codon gained:2
EEF1A1	CMX-G0000010487	1915	3189	2	Start codon gained:2
FMR1	CMX-	2332	3775	2	Drastic

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	Number of Variants detected	Variant Description (type and count)
	G0000031614				nonsynonymous:1; Start codon gained:1
GDF1	CMX- G0000027183	2657	4214	2	Drastic nonsynonymous:1; CNV gain:1
HK3	CMX- G0000009361	3101	4925	2	Drastic nonsynonymous:2
IGF2	CMX- G0000016702	3481	5466	2	CNV gain:2
ISG15	CMX- G0000000029	9636	4053	2	CNV gain:2
JMY	CMX- G0000008593	133746	28916	2	Drastic nonsynonymous:2
KL	CMX- G0000020228	9365	6344	2	Drastic nonsynonymous:2
MTHFR	CMX- G0000000213	4524	7436	2	Drastic nonsynonymous:1; Start codon gained:1
NLRP13	CMX- G0000028190	126204	22937	2	Drastic nonsynonymous:2
NLRP5	CMX- G0000028192	126206	21269	2	Drastic nonsynonymous:2
NOBOX	CMX- G0000012690	135935	22448	2	Drastic nonsynonymous:2
PRKRA	CMX- G0000004587	8575	9438	2	Drastic nonsynonymous:1; Nonsynonymous start:1
SDC3	CMX- G0000000574	9672	10660	2	Drastic nonsynonymous:2
TACC3	CMX- G0000006818	10460	11524	2	Drastic nonsynonymous:2
TLE6	CMX- G0000026639	79816	30788	2	CNV loss:2
ACVR1C	CMX- G0000004406	130399	18123	1	Drastic nonsynonymous:1
AHR	CMX- G0000011332	196	348	1	Start codon gained:1
APOA1	CMX- G0000018327	335	600	1	CNV gain:1
AURKA	CMX- G0000028967	6790	11393	1	Start codon gained:1

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	Number of Variants detected	Variant Description (type and count)
BMP15	CMX-G0000030783	9210	1068	1	CNV gain:1
BMP4	CMX-G0000021216	652	1071	1	Stop codon lost:1
C6orf221	CMX-G0000010478	154288	33699	1	Drastic nonsynonymous:1
CASP8	CMX-G0000004721	841	1509	1	CNV loss:1
CBS	CMX-G0000029408	875	1550	1	Drastic nonsynonymous:1
CDX2	CMX-G0000020191	1045	1806	1	Drastic nonsynonymous:1
CENPF	CMX-G0000002670	1063	1857	1	Drastic nonsynonymous:1
CGB	CMX-G0000027860	1082	1886	1	Start codon gained:1
CSF1	CMX-G0000001374	1435	2432	1	CNV loss:1
CSF2	CMX-G0000008885	1437	2434	1	CNV loss:1
DCTPP1	CMX-G0000023705	79077	28777	1	CNV gain:1
DNMT1	CMX-G0000026880	1786	2976	1	Drastic nonsynonymous:1
EFNA4	CMX-G0000001896	1945	3224	1	CNV loss:1
EFNB3	CMX-G0000024616	1949	3228	1	CNV gain:1
EIF3CL	CMX-G0000023621	728689	26347	1	CNV loss:1
EPHA5	CMX-G0000007213	2044	3389	1	CNV loss:1
EPHA7	CMX-G0000010603	2045	3390	1	CNV loss:1
EZH2	CMX-G0000012702	2146	3527	1	Drastic nonsynonymous:1
FOXL2	CMX-G0000006297	668	1092	1	Start codon gained:1
FOXP3	CMX-G0000030750	50943	6106	1	CNV gain:1
GALT	CMX-G0000014248	2592	4135	1	Splice site acceptor:1

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	Number of Variants detected	Variant Description (type and count)
GDF9	CMX-G0000008902	2661	4224	1	Start codon gained:1
GJA4	CMX-G0000000643	2701	4278	1	CNV gain:1
GJB3	CMX-G0000000642	2707	4285	1	CNV gain:1
GJB4	CMX-G0000000641	127534	4286	1	CNV gain:1
GJD3	CMX-G0000025169	125111	19147	1	CNV gain:1
GPC3	CMX-G0000031486	2719	4451	1	CNV gain:1
HSD17B2	CMX-G0000024260	3294	5211	1	Drastic nonsynonymous:1
IGFBPL1	CMX-G0000014341	347252	20081	1	CNV loss:1
KISS1	CMX-G0000002533	3814	6341	1	CNV gain:1
LHCGR	CMX-G0000003462	3973	6585	1	Drastic nonsynonymous:1
MAD1L1	CMX-G0000011200	8379	6762	1	Start codon gained:1
MAD2L1	CMX-G0000007650	4085	6763	1	Start codon gained:1
MB21D1	CMX-G0000010484	115004	21367	1	Drastic nonsynonymous:1
MCM8	CMX-G0000028433	84515	16147	1	Drastic nonsynonymous:1
MYC	CMX-G0000013826	4609	7553	1	Start codon gained:1
NLRP2	CMX-G0000028140	55655	22948	1	Start codon gained:1
NLRP4	CMX-G0000028189	147945	22943	1	Start codon gained:1
OAS1	CMX-G0000019838	4938	8086	1	Splice site acceptor:1
PADI3	CMX-G0000000342	51702	18337	1	CNV gain:1
PAEP	CMX-G0000015254	5047	8573	1	CNV gain:1
PLCB1	CMX-G0000028445	23236	15917	1	CNV gain:1

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	Number of Variants detected	Variant Description (type and count)
PMS2	CMX-G0000011251	5395	9122	1	Drastic nonsynonymous:1
POF1B	CMX-G0000031099	79983	13711	1	CNV gain:1
PRDM9	CMX-G0000008219	56979	13994	1	CNV loss:1
SEPHS2	CMX-G0000023707	22928	19686	1	CNV gain:1
SERPINA10	CMX-G0000021629	51156	15996	1	CNV gain:1
SIRT3	CMX-G0000016629	23410	14931	1	CNV loss:1
SPN	CMX-G0000023664	101929889	11249	1	CNV loss:1
TFPI	CMX-G0000004632	7035	11760	1	Drastic nonsynonymous:1
TGFB1H	CMX-G0000023757	7041	11767	1	CNV gain:1
TP63	CMX-G0000006674	8626	15979	1	Start codon gained:1
UBE3A	CMX-G0000022200	7337	12496	1	Start codon gained:1
UBL4B	CMX-G0000001378	164153	32309	1	CNV loss:1
UIMC1	CMX-G0000009362	51720	30298	1	Drastic nonsynonymous:1
VKORC1	CMX-G0000023741	79001	23663	1	CNV gain:1
ZP3	CMX-G0000011947	7784	13189	1	Start codon gained:1

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 2 below. In Table 2, HGNC

5 (<http://www.genenames.org/>) reference numbers are provided when available.

Table 2 below depicts another possible gene ranking scheme for the relative infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. Table 2 contains the 10 genes, listed in order from most to least statistically significant, that were determined to be statistically significantly correlated with

infertility risk in a study of unexplained female infertility based on variants detected in the coding regions of these genes. P-values < .025 are considered statistically significant, and all other fertility genes did not pass the significance test for inclusion and ranking in this list. For the coding level analysis, we first compute a coding variant score for the coding regions for each individual/gene. The coding variant score represents the variability of the gene at coding regions in an individual and is computed as the sum of the proportion of variant locations within the coding regions of that gene for that individual. A series of linear regression models are fit, where the outcome variable is the coding variant score for a given gene, and the independent variables are group (infertile vs control) and principal component derived ethnicity (continuous). The p-value for group is used for statistical inference. The model is fit once for each gene.

Table 2: Fertility genes demonstrating statistical significance at the gene coding region level for infertility risk ranked based on p-values, observed in a study of unexplained female infertility.

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	P-value
ZP4	CMX- G0000002903	57829	15770	5.17E-10
UIMC1	CMX- G0000009362	51720	30298	0.001401803
PADI6	CMX- G0000000344	353238	20449	0.003420271
ZP1	CMX- G0000017558	22917	13187	0.003845858
MDM2	CMX- G0000019503	4193	6973	0.009323844
PRKRA	CMX- G0000004587	8575	9438	0.009832035
PMS2	CMX- G0000011251	5395	9122	0.015453858
TGFB1	CMX- G0000027588	7040	11766	0.018576967
ESR2	CMX- G0000021326	2100	3468	0.022661688
PRDM1	CMX- G0000010653	639	9346	0.024522163

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 3 below. In Table 3, HGNC (<http://www.genenames.org/>) reference numbers are provided when available.

Table 3 below depicts another possible gene ranking scheme for the relative infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. Table 3 contains the 11 genes, listed in order from most to least statistically significant, that were determined to be statistically significantly correlated with infertility risk in a study of unexplained female infertility based on variants detected in the coding, non-coding, and conserved upstream and downstream regions of the fertility gene. P-values < .025 are considered statistically significant, and all other fertility genes did not fit the pass the significance test for inclusion and ranking in this list. For the gene level analysis, we first compute a gene variant score for the entire transcript and flanking evolutionarily conserved regions for each individual/ gene. The gene variant score represents the variability of the gene in an individual and is computed as the sum of the proportion of variant locations within that gene and its evolutionarily conserved regions flanking the gene for that individual. A series of linear regression models are fit, where the outcome variable is the gene variant score for a given gene, and the independent variables are group (infertile vs control) and principal component derived ethnicity (continuous). The p-value for group is used for statistical inference. The model is fit once for each gene.

Table 3: Fertility genes demonstrating statistical significance at the entire gene level for infertility risk ranked based on p-values, observed in a study of unexplained female infertility.

Gene	Celmatix Gene ID	Entrez ID	HGNC ID	P-value
PADI6	CMX-G0000000344	353238	20449	0.00079599
CGB	CMX-G0000027860	1082	1886	0.000983714
PMS2	CMX-G0000011251	5395	9122	0.001500248
ESR2	CMX-G0000021326	2100	3468	0.004733531
UIMC1	CMX-G0000009362	51720	30298	0.005170633

ZP1	CMX-G0000017558	22917	13187	0.00852914
MDM2	CMX-G0000019503	4193	6973	0.009794758
BRCA2	CMX-G0000020222	675	1101	0.019744499
TGFB1	CMX-G0000027588	7040	11766	0.020358934
CDKN1C	CMX-G0000016717	1028	1786	0.022605239
TAF4B	CMX-G0000026229	6875	11538	0.024673723

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 4 below. In Table 4, HGNC

5 (<http://www.genenames.org/>) reference numbers are provided when available.

Table 4 below depicts another possible gene ranking scheme for the relative infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. Table 4 contains the top ranked 100 fertility genes, listed in order from most to least likely for variants in that gene to affect fertility. Genes are ranked according to a Celmatix Fertilome™Score, G1Version2, that reflects the likelihood a gene is involved in fertility or reproduction. This score is computed using a database of mined and curated data, containing attributes for each gene in the genome (See Figures 5 and 6). These attributes include: diseases and disorders related to infertility, molecular pathways, molecular interactions, gene clusters, mouse phenotypes associated with each gene, gene expression data in reproductive tissues, proteomics data in oocytes, and accrued information from scientific publications through text-mining.

The process for ranking fertility-related attributes of a gene or genetic region (locus) to obtain an infertility score is called the SESMe algorithm. The SESMe algorithm is applied to a database of features and attributes that might make a particular gene important for fertility. The algorithm assigns a score and a relative weight to each feature then ranks genetic regions from most to least important (or vice versa) by weighting features and attributes associated with that genetic region. For example, a score is assigned to a gene by compiling the combined weighted values of attributes associated with that gene. After each gene is scored based on its weighted attributes, the genetic loci can be ranked in order of importance in accordance with their score. The weighted value for each infertility attribute may be scaled in any manner including and not

limited to assigning a positive or negative integer to reflect the significance or severity of the attribute to infertility.

In certain embodiments, the weighted value for gene infertility attributes may be on a scale from -10 to +10. A +10 may indicate that an attribute of a gene being scored is highly associated with infertility because that attribute is prevalently found in infertile patient populations. A +4 may represent an attribute that is a latent infertility marker, meaning it will not cause infertility on its own, but may lead to infertility upon influence of external factors such as aging and smoking. Whereas +2 may represent an attribute found in some infertile patients but nothing directly relates the attribute to infertility. A zero on the scale may include an attribute not yet known to have any effect or any negative effect towards infertility. A -10 may include an attribute shown not to affect infertility whatsoever. Further, embodiments provide for the weighted scale to include a +1 for attributes that are commonly found in infertile patient populations, 0.5 for attributes similar to those found in infertile patient populations, and 0 for attributes without a causal link to infertility.

In addition, weighted values for attributes may be normalized based on the known significance of that attribute towards infertility. For example and in certain embodiments, when scoring attributes of a particular gene, each attribute may be assigned a 0 if the attribute is absent and a 1 if the attribute is present. The attributes may then be normalized based on the infertility significance of that attribute. For example, if the attribute is a genetic mutation known to be associated with infertility, then that attribute may be normalized by a factor of 5. In another example, if the attribute is a signaling pathway defect sometimes associated with infertility, then that attribute may be normalized by a factor of 2.

Table 4, provided below, lists 100 Human Fertility Genes that were ranked by weighing attributes associated with the gene in accordance with methods of the invention.

Table 4: List of Top 100 Human Fertility Genes based on the Fertilome™ Score, G1Version2.

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
1	C6orf221	CMX-G00000010478	154288	33699	15

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
2	NLRP5	CMX-G0000028192	126206	21269	15
3	ZP3	CMX-G0000011947	7784	13189	12.93
4	FIGLA	CMX-G0000003616	344018	24669	12
5	PADI6	CMX-G0000000344	353238	20449	12
6	DNMT1	CMX-G0000026880	1786	2976	11.67
7	ZP2	CMX-G0000023549	7783	13188	11.67
8	FSHR	CMX-G0000003464	2492	3969	11.37
9	OOEP	CMX-G0000010479	441161	21382	11
10	FOXO3	CMX-G0000010672	2309	3821	10.39
11	ACVR1B	CMX-G0000019186	91	172	10.14
12	CGA	CMX-G0000010560	1081	1885	10.04
13	INHA	CMX-G0000004914	3623	6065	10.02
14	LHCGR	CMX-G0000003462	3973	6585	10.01
15	DPPA3	CMX-G0000018719	359787	19199	10
16	KDM1B	CMX-G0000009642	221656	21577	10
17	NOBOX	CMX-G0000012690	135935	22448	10
18	NPM2	CMX-G0000013114	10361	7930	10
19	ESR1	CMX-G0000011002	2099	3467	9.91
20	AURKA	CMX-G0000028967	6790	11393	9.84
21	BRCA2	CMX-G0000020222	675	1101	9.75
22	WT1	CMX-G0000017126	7490	12796	9.53
23	CBS	CMX-G0000029408	875	1550	9.49
24	CDKN1C	CMX-G0000016717	1028	1786	9.37
25	IGF1	CMX-G0000019714	3479	5464	9.35
26	HAND2	CMX-G0000007954	9464	4808	9.17
27	GDF9	CMX-G0000008902	2661	4224	9
28	MAD2L1	CMX-G0000007650	4085	6763	9
29	ZAR1	CMX-G0000007128	326340	20436	9
30	FOXL2	CMX-G0000006297	668	1092	8.88
31	BARD1	CMX-G0000004834	580	952	8.54
32	FMN2	CMX-G0000002910	56776	14074	8.4
33	TACC3	CMX-G0000006818	10460	11524	8.39
34	MYC	CMX-G0000013826	4609	7553	8.25
35	IL11RA	CMX-G0000014249	3590	5967	7.9
36	MCM8	CMX-G0000028433	84515	16147	7.85
37	LHB	CMX-G0000027859	3972	6584	7.82
38	TAF4B	CMX-G0000026229	6875	11538	7.68

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
39	USP9X	CMX-G0000030612	8239	12632	7.67
40	PRLR	CMX-G0000008271	5618	9446	7.58
41	HSF1	CMX-G0000013948	3297	5224	7.35
42	FSHB	CMX-G0000017113	2488	3964	7.33
43	ZP1	CMX-G0000017558	22917	13187	7.29
44	MDM2	CMX-G0000019503	4193	6973	7.27
45	BMP15	CMX-G0000030783	9210	1068	7.25
46	GPC3	CMX-G0000031486	2719	4451	7.11
47	PRDM1	CMX-G0000010653	639	9346	7.05
48	FST	CMX-G0000008371	10468	3971	7
49	EZH2	CMX-G0000012702	2146	3527	6.91
50	SMAD2	CMX-G0000026329	4087	6768	6.89
51	NODAL	CMX-G0000015959	4838	7865	6.88
52	ACVR1	CMX-G0000004407	90	171	6.81
53	HSD17B12	CMX-G0000017190	51144	18646	6.71
54	BRCA1	CMX-G0000025305	672	1100	6.67
55	DICER1	CMX-G0000021645	23405	17098	6.53
56	ESR2	CMX-G0000021326	2100	3468	6.47
57	MDM4	CMX-G0000002542	4194	6974	6.42
58	AR	CMX-G0000030935	367	644	6.41
59	SCARB1	CMX-G0000019991	949	1664	6.39
60	CDKN1B	CMX-G0000018846	1027	1785	6.25
61	TP53	CMX-G0000024614	7157	11998	6.23
62	NOG	CMX-G0000025542	9241	7866	6.22
63	IL6ST	CMX-G0000008398	3572	6021	6.13
64	DAZL	CMX-G0000005296	1618	2685	6
65	NLRP11	CMX-G0000028188	204801	22945	6
66	NLRP13	CMX-G0000028190	126204	22937	6
67	NLRP8	CMX-G0000028191	126205	22940	6
68	NLRP9	CMX-G0000028184	338321	22941	6
69	ZFX	CMX-G0000030503	7543	12869	5.67
70	TFPI	CMX-G0000004632	7035	11760	5.36
71	HSD17B7	CMX-G0000002148	51478	5215	5.32
72	TP63	CMX-G0000006674	8626	15979	5.28
73	NR5A1	CMX-G0000015051	2516	7983	5.24
74	BMP7	CMX-G0000028985	655	1074	5.09
75	CGB	CMX-G0000027860	1082	1886	5

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
76	CGB5	CMX-G0000027866	93659	16452	5
77	DDX43	CMX-G0000010483	55510	18677	5
78	FMR1	CMX-G0000031614	2332	3775	5
79	LIN28B	CMX-G0000010647	389421	32207	5
80	NLRP14	CMX-G0000016919	338323	22939	5
81	NLRP4	CMX-G0000028189	147945	22943	5
82	NLRP7	CMX-G0000028139	199713	22947	5
83	PROK1	CMX-G0000001385	84432	18454	5
84	SPIN1	CMX-G0000014689	10927	11243	5
85	TFPI2	CMX-G0000012044	7980	11761	5
86	ZP4	CMX-G0000002903	57829	15770	5
87	ESRRB	CMX-G0000021489	2103	3473	4.8
88	UBE3A	CMX-G0000022200	7337	12496	4.76
89	SUZ12	CMX-G0000025003	23512	17101	4.73
90	XIST	CMX-G0000031023	7503	12810	4.7
91	ATM	CMX-G0000018234	472	795	4.62
92	AURKB	CMX-G0000024639	9212	11390	4.55
93	STK3	CMX-G0000013673	6788	11406	4.52
94	POLG	CMX-G0000023009	5428	9179	4.51
95	CDX2	CMX-G0000020191	1045	1806	4.46
96	TP73	CMX-G0000000110	7161	12003	4.43
97	MTOR	CMX-G0000000201	2475	3942	4.42
98	AHR	CMX-G0000011332	196	348	4.41
99	LIF	CMX-G0000029949	3976	6596	4.38
100	PRKRA	CMX-G0000004587	8575	9438	4.38

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 5 below. In Table 5, HGNC

5 (<http://www.genenames.org/>) reference numbers are provided when available.

Table 5 below depicts another possible gene ranking scheme for the relative infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. Table 5 contains the top ranked 100 fertility genes, listed in order from most to least likely for variants in that gene to affect fertility. Genetic loci are ranked according to a Celmatix Fertilome™ Score, G1Version3, that reflects the likelihood a gene is

10

involved in fertility or reproduction. This score is computed using a database of mined and curated data, containing attributes for each gene in the genome (See Figures 5 and 6). These attributes include: diseases and disorders related to infertility, molecular pathways, molecular interactions, gene clusters, mouse phenotypes associated with each gene, gene expression data in reproductive tissues, proteomics data in oocytes, and accrued information from scientific publications through text-mining. The Celmatix Fertilome™Score, G1Version3 differs from G1Version2 (Table 4) because it contains more fertility genes as an input for the score calculation.

Table 5: List of Top 100 Human Fertility Genes based on the Fertilome™Score, G1Version3.

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
1	C6orf221	CMX-G0000010478	154288	33699	15
2	NLRP5	CMX-G0000028192	126206	21269	15
3	TCL1A	CMX-G0000021654	8115	11648	14
4	ZP3	CMX-G0000011947	7784	13189	12.93
5	FIGLA	CMX-G0000003616	344018	24669	12
6	PADI6	CMX-G0000000344	353238	20449	12
7	RSPO1	CMX-G0000000687	284654	21679	12
8	EPHA1	CMX-G0000012650	2041	3385	11.82
9	DNMT1	CMX-G0000026880	1786	2976	11.67
10	ZP2	CMX-G0000023549	7783	13188	11.67
11	MOS	CMX-G0000013392	4342	7199	11.5
12	FSHR	CMX-G0000003464	2492	3969	11.37
13	OOEP	CMX-G0000010479	441161	21382	11

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
14	CUL1	CMX-G0000012701	8454	2551	10.67
15	HSP90B1	CMX-G0000019724	7184	12028	10.57
16	FOXO3	CMX-G0000010672	2309	3821	10.39
17	KISS1	CMX-G0000002533	3814	6341	10.21
18	ACVR1B	CMX-G0000019186	91	172	10.14
19	CGA	CMX-G0000010560	1081	1885	10.04
20	INHA	CMX-G0000004914	3623	6065	10.02
21	LHCGR	CMX-G0000003462	3973	6585	10.01
22	DPPA3	CMX-G0000018719	359787	19199	10
23	KDM1B	CMX-G0000009642	221656	21577	10
24	NOBOX	CMX-G0000012690	135935	22448	10
25	NPM2	CMX-G0000013114	10361	7930	10
26	PRMT3	CMX-G0000017073	10196	30163	10
27	GJA4	CMX-G0000000643	2701	4278	9.92
28	ESR1	CMX-G0000011002	2099	3467	9.91
29	SFRP4	CMX-G0000011506	6424	10778	9.89
30	AURKA	CMX-G0000028967	6790	11393	9.84
31	BRCA2	CMX-G0000020222	675	1101	9.75
32	WT1	CMX-G0000017126	7490	12796	9.53
33	CBS	CMX-G0000029408	875	1550	9.49
34	CDKN1C	CMX-G0000016717	1028	1786	9.37
35	IGF1	CMX-	3479	5464	9.35

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilo™ Score
		G0000019714			
36	PLCB1	CMX-G0000028445	23236	15917	9.33
37	CEP290	CMX-G0000019604	80184	29021	9.3
38	MSH5	CMX-G0000010000	4439	7328	9.29
39	HAND2	CMX-G0000007954	9464	4808	9.17
40	GDF9	CMX-G0000008902	2661	4224	9
41	MAD2L1	CMX-G0000007650	4085	6763	9
42	TNFAIP6	CMX-G0000004377	7130	11898	9
43	ZAR1	CMX-G0000007128	326340	20436	9
44	FOXL2	CMX-G0000006297	668	1092	8.88
45	PCNA	CMX-G0000028417	5111	8729	8.78
46	YBX2	CMX-G0000024578	51087	17948	8.57
47	BARD1	CMX-G0000004834	580	952	8.54
48	AMBP	CMX-G0000014963	259	453	8.4
49	FMN2	CMX-G0000002910	56776	14074	8.4
50	NCOA2	CMX-G0000013477	10499	7669	8.4
51	TEX12	CMX-G0000018279	56158	11734	8.4
52	TACC3	CMX-G0000006818	10460	11524	8.39
53	PGR	CMX-G0000018173	5241	8910	8.37
54	FANCC	CMX-G0000014774	2176	3584	8.25
55	MYC	CMX-G0000013826	4609	7553	8.25
56	FGF8	CMX-G0000016316	2253	3686	8.23

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix FertiloMe™ Score
57	SMAD5	CMX- G0000008943	4090	6771	8.12
58	CCS	CMX- G0000017793	9973	1613	8
59	MSH4	CMX- G0000001108	4438	7327	8
60	SPO11	CMX- G0000028986	23626	11250	8
61	SYCE1	CMX- G0000016602	93426	28852	8
62	SYCP1	CMX- G0000001457	6847	11487	8
63	TFAP2C	CMX- G0000028982	7022	11744	8
64	WNT7A	CMX- G0000005260	7476	12786	7.96
65	IL11RA	CMX- G0000014249	3590	5967	7.9
66	MCM8	CMX- G0000028433	84515	16147	7.85
67	SYCP2	CMX- G0000029020	10388	11490	7.85
68	INHBA	CMX- G0000011550	3624	6066	7.83
69	MGAT1	CMX- G0000009451	4245	7044	7.83
70	LHB	CMX- G0000027859	3972	6584	7.82
71	CYP19A1	CMX- G0000022537	1588	2594	7.74
72	GGT1	CMX- G0000029874	2678	4250	7.71
73	TAF4B	CMX- G0000026229	6875	11538	7.68
74	SMC1B	CMX- G0000030247	27127	11112	7.67
75	USP9X	CMX- G0000030612	8239	12632	7.67
76	PRLR	CMX- G0000008271	5618	9446	7.58
77	DNMT3B	CMX- G0000028640	1789	2979	7.54
78	SOD1	CMX-	6647	11179	7.54

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilo™ Score
		G0000029263			
79	SH2B1	CMX-G0000023639	25970	30417	7.5
80	HOXA11	CMX-G0000011417	3207	5101	7.48
81	UBB	CMX-G0000024729	7314	12463	7.43
82	HSF1	CMX-G0000013948	3297	5224	7.35
83	CYP17A1	CMX-G0000016340	1586	2593	7.33
84	FSHB	CMX-G0000017113	2488	3964	7.33
85	SYCP3	CMX-G0000019706	50511	18130	7.33
86	NOS3	CMX-G0000012751	4846	7876	7.31
87	ZP1	CMX-G0000017558	22917	13187	7.29
88	GNRHR	CMX-G0000007221	2798	4421	7.27
89	MDM2	CMX-G0000019503	4193	6973	7.27
90	BMP15	CMX-G0000030783	9210	1068	7.25
91	KDM1A	CMX-G0000000422	23028	29079	7.25
92	MDK	CMX-G0000017221	4192	6972	7.21
93	MSX2	CMX-G0000009331	4488	7392	7.21
94	CTNNB1	CMX-G0000005462	1499	2514	7.2
95	NRIP1	CMX-G0000029160	8204	8001	7.2
96	UBC	CMX-G0000019992	7316	12468	7.2
97	FKBP4	CMX-G0000018615	2288	3720	7.19
98	MLH3	CMX-G0000021470	27030	7128	7.14
99	MSX1	CMX-G0000006873	4487	7391	7.13

Rank	Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID	Celmatix Fertilome™ Score
100	GPC3	CMX-G0000031486	2719	4451	7.11

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 6 below. In Table 5, HGNC (<http://www.genenames.org/>) reference numbers are provided when available.

Table 6 below depicts another possible gene ranking scheme for the relative infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. Table 6 contains the top ranked fertility genes based on a comparison of how often the gene appears in one of the lists above (Tables 1-5). This list represents the top 20 genetic regions with utility for diagnosing female infertility, subfertility, or premature decline in fertility. These targets were identified using a compendium of factors: 1) Carrying statistically significant genetic mutations at the coding level in a pilot study, 2) Carrying statistically significant genetic mutations at the coding level in a pilot study, 3) Carrying genetic variations in our pilot study that impact the biochemical properties of the gene, 4) Highly ranked in our Celmatix Fertilome™ Score system, that reflects the likelihood a gene is involved in fertility or reproduction.

Table 6: List of the Top 20 Fertility Genes (arranged in alphabetical order)

Gene Symbol	Celmatix Gene ID	Entrez Gene ID	HGNC Gene ID
BARD1	CMX-G0000004834	580	952
C6orf221	CMX-G0000010478	154288	33699
DNMT1	CMX-G0000026880	1786	2976

	CMX-		
FMR1	G0000031614	2332	3775
	CMX-		
FOXO3	G0000010672	2309	3821
	CMX-		
MUC4	G0000006719	4585	7514
	CMX-		
NLRP11	G0000028188	204801	22945
	CMX-		
NLRP14	G0000016919	338323	22939
	CMX-		
NLRP5	G0000028192	126206	21269
	CMX-		
NLRP8	G0000028191	126205	22940
	CMX-		
NPM2	G0000013114	10361	7930
	CMX-		
PADI6	G0000000344	353238	20449
	CMX-		
PMS2	G0000011251	5395	9122
	CMX-		
SCARB1	G0000019991	949	1664
	CMX-		
SPIN1	G0000014689	10927	11243
	CMX-		
TACC3	G0000006818	10460	11524
	CMX-		
ZP1	G0000017558	22917	13187
	CMX-		
ZP2	G0000023549	7783	13188
ZP3	CMX-	7784	13189

G0000011947			
CMX-			
ZP4	G0000002903	57829	15770

In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility selected from the genes shown in Table 7 below. In Table 7, HGNC
 5 (<http://www.genenames.org/>) reference numbers are provided when available.

Table 7 below depicts all of the biologically and/or statistically significant variants detected in the genes depicted in Table 6 in a genetic study of female infertility. Genetic variants considered to be biologically significant include mutations that result in a change: 1) to a different amino acid predicted to alter the folding and/or structure of the encoded protein, 2) to a different
 10 amino acid occurring at a highly evolutionarily conserved site, 3) that introduces a premature stop termination signal, 4) that causes a stop termination signal to be lost, 5) that introduces a new start codon, 6) that causes a start codon to be lost, 7) that disrupts a splicing signal, 8) that alters the reading frame or 9) that alters the dosage of encoded protein or RNA. All genetic variants detected from resequencing exclude sites at the single nucleotide level where the variant
 15 allele is detected in only one chromosome (singletons) and sites sequenced in only one individual. Structural variants impacting biological function are also reported. Using these criteria applied to targeted re-sequencing data from a study of infertile females, we detected 490 variants, of which 379 are listed in Table 7.

For the statistically significant variant level analysis, a series of
 20 logistic regression models are fit, where the outcome variable is the binary indicator of variant status for a given location, and the independent variables are group (infertile vs. control) and principal component-derived ethnicity (continuous). The p-value and odds ratio for group are used for statistical inference. The model is fit once for each location. P-values < .001 are considered statistically significant. We performed a SNP association study by targeted re-
 25 sequencing and identified a total of 147 SNPs significantly associated with female infertility (of which 52 are reported in Table 7). Each variant was classified as novel or known. Novel sites are excluded from the p-value computation. For known variants, we apply a series of logistic regression models where the outcome variable is the binary indicator of variant status for a given

location, and the independent variables are group (infertile vs. control) and principal component-derived ethnicity (continuous). The p-value and odds ratio for group are used for statistical inference. P-values less than .001 were considered significant. Position refers to NCBI Build 37. Alleles are reported on the forward strand. Ref = Reference allele, Alt = Variant allele.

5

Table 7: List of Biologically and Statistically Significant Genetic Variants Most Useful for Predicting Infertility Risk in Humans (arranged in alphabetical order by gene name)

Gene Symbol	Celmatix Gene ID	Celmatix Variant ID	Location	Ref	Alt	Impact	P-value
APOA1	CMX-G0000018327	CMX-V1388879	chr11:1125539-69-126265772	NA	CN V gain	APOA1 (3 exons)	NA
ASCL2	CMX-G0000016707	CMX-V1067111	chr11:2234334-2298706	NA	CN V gain	ASCL2 (1 exon)	NA
BARD1	CMX-G0000004834	CMX-V9083698	chr2:215674224	G	A	Drastic nonsynonymous	NA
BARD1	CMX-G0000004834	CMX-V9083699	chr2:215595645	C	T	Start codon lost	NA
BARD1	CMX-G0000004834	CMX-V9083700	chr2:215674323	C	G	Start codon gained	NA
BARD1	CMX-G0000004834	CMX-SV00001	chr2:215645502	GTGGTG AAGAAC ATTCAG GCAA	G	Codon deletion	NA
BARD1	CMX-G0000004834	CMX-V9084177	chr2:215742204	G	T	NA	6.77E-05
BMP15	CMX-G0000030783	CMX-V1250077	chrX:5063996-9-50981841	NA	CN V gain	BMP15 (2 exons)	NA
BMP6	CMX-G0000009564	CMX-V1247770	chr6:7726514-7727614	NA	CN V loss	BMP6 (1 exon)	NA
BMP6	CMX-G0000009564	CMX-V1166409	chr6:7724859-7728905	NA	CN V loss	BMP6 (1 exon)	NA
C6orf221	CMX-G0000010478	CMX-V9083706	chr6:74073531	C	G	Drastic nonsynonymous	NA
CASP8	CMX-	CMX-	chr2:20185112	NA	CN	CASP8 (2	NA

	G00000047 21	V1843349	9-203110758		V loss	exons)	
CSF1, BL4B	CMX- G00000013 74,CMX- G00000013 78	CMX- V1667025	chr1:11044146 5-110831379	NA	CN V loss	CSF1 (4 exons), UBL4B (1 exon)	NA
	CMX- G00000088 85	CMX- V1456214	chr5:12832021 8-131440732	NA	CN V loss	CSF2 (4 exons)	NA
CYP11 B1	CMX- G00000138 88	CMX- V1957973	chr8:14395181 3-143958440	NA	CN V gain	CYP11B1 (4 exons)	NA
CYP11 B1	CMX- G00000138 88	CMX- V1609269	chr8:14395340 3-143991713	NA	CN V gain	CYP11B1 (4 exons)	NA
DCTPP 1,SEPH S2,TGF B111,V KORC1	CMX- G00000237 05,CMX- G00000237 07,CMX- G00000237 57,CMX- G00000237 41	CMX- V1070550	chr16:3034768 9-31632796	NA	CN V gain	DCTPP1 (1 exon), SEPHS2 (1 exon), TGFB111 (3 exons), VKORC1 (1 exon)	NA
	CMX- G00000268 80	CMX- V9083720	chr19:1029118 1	T	C	Drastic nonsynonymo us	NA
ECHS1	CMX- G00000165 94	CMX- V1101514	chr10:1350870 81-135243330	NA	CN V gain	ECHS1 (8 exons)	NA
ECHS1	CMX- G00000165 94	CMX- V1131837	chr10:1350888 39-135243616	NA	CN V loss	ECHS1 (8 exons)	NA
ECHS1	CMX- G00000165 94	CMX- V1335364	chr10:1350879 62-135243616	NA	CN V gain	ECHS1 (8 exons)	NA
EFNA4	CMX- G00000018 96	CMX- V1267541	chr1:15435457 6-155066744	NA	CN V loss	EFNA4 (4 exons)	NA
EFNB3	CMX- G00000246 16	CMX- V1295730	chr17:7135639 -7702377	NA	CN V gain	EFNB3 (5 exons)	NA
EIF3C L	CMX- G00000236 21	CMX- V1992389	chr16:2819703 2-28410526	NA	CN V loss	EIF3CL (13 exons)	NA
EPHA5	CMX- G00000072 13	CMX- V1585842	chr4:66114884 -66870165	NA	CN V loss	EPHA5 (17 exons)	NA
EPHA7	CMX- G00000106	CMX- V1939194	chr6:94015504 -95364976	NA	CN V	EPHA7 (3 exons)	NA

	03				loss		
	CMX-				CN		
	G00000004	CMX-	chr1:22906197		V	EPHA8 (1	
EPHA8	15	V1493926	-22914076	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22905731		V	EPHA8 (2	
EPHA8	15	V1680494	-22915711	NA	loss	exons)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22904786		V	EPHA8 (2	
EPHA8	15	V1333389	-22915711	NA	loss	exons)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22906271		V	EPHA8 (2	
EPHA8	15	V1750787	-22915711	NA	loss	exons)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22906197		V	EPHA8 (1	
EPHA8	15	V1102470	-22915047	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22905731		V	EPHA8 (1	
EPHA8	15	V1356293	-22915352	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22905731		V	EPHA8 (1	
EPHA8	15	V1845595	-22913963	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22906526		V	EPHA8 (1	
EPHA8	15	V1973671	-22913011	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22905731		V	EPHA8 (2	
EPHA8	15	V1086453	-22916983	NA	loss	exons)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22904856		V	EPHA8 (1	
EPHA8	15	V1138079	-22913700	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22904786		V	EPHA8 (1	
EPHA8	15	V1957426	-22914210	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22906197		V	EPHA8 (1	
EPHA8	15	V1635641	-22915352	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22905731		V	EPHA8 (1	
EPHA8	15	V1387198	-22914256	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22906271		V	EPHA8 (1	
EPHA8	15	V1481340	-22913750	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22904856		V	EPHA8 (1	
EPHA8	15	V1077862	-22913963	NA	loss	exon)	NA
	CMX-				CN		
	G00000004	CMX-	chr1:22904064		V	EPHA8 (1	
EPHA8	15	V1288029	-22914256	NA	loss	exon)	NA
EPHA8	CMX-	CMX-	chr1:22906395	NA	CN	EPHA8 (1	NA

	G00000004 15	V1098423	-22913750		V loss	exon)	
	CMX- G00000004 15	CMX- V1825294	chr1:22906271 -22914210	NA	CN V loss	EPHA8 (1 exon)	NA
EPHA8	CMX- G00000004 15	CMX- V1672255	chr1:22906271 -22915161	NA	CN V loss	EPHA8 (1 exon)	NA
	CMX- G00000004 15	CMX- V1740010	chr1:22906271 -22914076	NA	CN V loss	EPHA8 (1 exon)	NA
EPHA8	CMX- G00000004 15	CMX- V1757241	chr1:22904856 -22915352	NA	CN V loss	EPHA8 (1 exon)	NA
	CMX- G00000004 15	CMX- V1080982	chr1:22906322 -22914695	NA	CN V loss	EPHA8 (1 exon)	NA
EPHA8	CMX- G00000004 15	CMX- V1506728	chr1:22905731 -22913502	NA	CN V loss	EPHA8 (1 exon)	NA
	CMX- G00000163 16	CMX- V1202186	chr10:1035244 44-103533748	NA	CN V gain	FGF8 (2 exons)	NA
FGF8	CMX- G00000163 16	CMX- V1242750	chr10:1035247 14-103532892	NA	CN V gain	FGF8 (2 exons)	NA
	CMX- G00000163 16	CMX- V1059642	chr10:1035200 69-103531134	NA	CN V gain	FGF8 (1 exon)	NA
FGF8	CMX- G00000163 16	CMX- V1478224	chr10:1035250 82-103536399	NA	CN V gain	FGF8 (6 exons)	NA
	CMX- G00000316 14	CMX- V9083727	chrX:1470102 63	A	C	Drastic nonsynonymo us	NA
FMR1	CMX- G00000316 14	CMX- V9083728	chrX:1470149 60	C	T	Start codon gained	NA
	CMX- G00000316 14	CMX- V9084252	chrX:1461264 83	G	A	NA	0.0001 98744
FMR1	CMX- G00000316 14	CMX- V9084253	chrX:1461539 70	C	T	NA	1.92E- 05
	CMX- G00000316 14	CMX- V9084254	chrX:1461958 65	A	G	NA	0.0003 71198
FMR1	CMX- G00000316 14	CMX- V9084255	chrX:1462215 14	C	T	NA	0.0002 92157

FMR1	CMX- G00000316 14	CMX- V9084256	chrX:1462477 40	T	A	NA	0.0001 997
FMR1	CMX- G00000316 14	CMX- V9084257	chrX:1462552 13	G	A	NA	0.0001 85975
FMR1	CMX- G00000316 14	CMX- V9084258	chrX:1464063 19	A	G	NA	0.0002 62855
FMR1	CMX- G00000316 14	CMX- V9084259	chrX:1469949 16	A	G	NA	0.0008 16693
FMR1	CMX- G00000316 14	CMX- V9084260	chrX:1470029 92	T	G	NA	0.0008 10806
FMR1	CMX- G00000316 14	CMX- V9084261	chrX:1470033 39	A	G	NA	0.0008 10806
FMR1	CMX- G00000316 14	CMX- V9084262	chrX:1470037 94	T	C	NA	0.0008 10806
FMR1	CMX- G00000316 14	CMX- V9084263	chrX:1470245 58	A	T	NA	0.0006 41561
FMR1	CMX- G00000316 14	CMX- V9084264	chrX:1473725 28	G	C	NA	0.0006 33948
FMR1	CMX- G00000316 14	CMX- V9084265	chrX:1473978 06	A	G	NA	0.0008 13685
FMR1	CMX- G00000316 14	CMX- V9084266	chrX:1474376 83	A	G	NA	0.0007 84981
FMR1	CMX- G00000316 14	CMX- V9084267	chrX:1474496 73	T	C	NA	0.0004 01568
FMR1	CMX- G00000316 14	CMX- V9084268	chrX:1474548 32	G	A	NA	0.0009 65078
FMR1	CMX- G00000316 14	CMX- V9084269	chrX:1474782 74	G	T	NA	0.0006 46517
FMR1	CMX- G00000316 14	CMX- V9084270	chrX:1474798 61	A	C	NA	0.0006 46517
FMR1	CMX- G00000316 14	CMX- V9084271	chrX:1474802 74	A	G	NA	0.0006 46517
FMR1	CMX- G00000316 14	CMX- V9084272	chrX:1474818 91	T	C	NA	0.0006 46517

FMR1	CMX- G00000316 14	CMX- V9084273	chrX:1474826 03	A	G	NA	0.0005 64877
FMR1	CMX- G00000316 14	CMX- V9084274	chrX:1474826 30	A	G	NA	0.0004 58631
FOXO3	CMX- G00000106 72	CMX- V9084196	chr6:10885610 8	C	T	NA	0.0002 32121
FOXO3	CMX- G00000106 72	CMX- V9084197	chr6:10914969 3	G	C	NA	0.0003 44433
FOXO3	CMX- G00000106 72	CMX- V9084195	chr6:10885336 1	T	A	NA	0.0001 76018
FOXO3	CMX- G00000106 72	CMX- V9084198	chr6:10915578 9	G	T	NA	0.0006 41107
FOXO3	CMX- G00000106 72	CMX- V1295244	chr6:10898514 8-108989762	NA	CN V gain	FOXO3 (1 exon)	NA
FOXO3	CMX- G00000106 72	CMX- V1963522	chr6:10898550 7-108989056	NA	CN V gain	FOXO3 (1 exon)	NA
FOXO3	CMX- G00000106 72	CMX- V1616823	chr6:10898493 0-108989762	NA	CN V gain	FOXO3 (1 exon)	NA
FOXP3	CMX- G00000307 50	CMX- V1008919	chrX:4889022 1-49257528	NA	CN V gain	FOXP3 (9 exons)	NA
GDF1	CMX- G00000271 83	CMX- V1625432	chr19:1887218 5-19535389	NA	CN V gain	GDF1 (2 exons)	NA
GJA4,G JB3,GJ B4	CMX- G00000006 43,CMX- G00000006 42,CMX- G00000006 41	CMX- V1706868	chr1:35000925 -37866010	NA	CN V gain	GJA4 (1 exon), GJB3 (1 exon), GJB4 (1 exon)	NA
GJD3	CMX- G00000251 69	CMX- V1132225	chr17:3795254 1-38532715	NA	CN V gain	GJD3 (1 exon)	NA
GPC3	CMX- G00000314 86	CMX- V1515961	chrX:1326139 06-132779666	NA	CN V gain	GPC3 (1 exon)	NA
IGF2	CMX- G00000167 02	CMX- V1454080	chr11:2127129 -2173473	NA	CN V gain	IGF2 (3 exons)	NA
IGF2	CMX- G00000167	CMX- V1542559	chr11:2110901 -2173938	NA	CN V	IGF2 (3 exons)	NA

	02				gain		
IGFBP L1	CMX- G00000143 41	CMX- V1435664	chr9:35776310 -38419649	NA	CN V loss	IGFBPL1 (3 exons)	NA
ISG15	CMX- G00000000 29	CMX- V1111642	chr1:940142- 1016233	NA	CN V gain	ISG15 (2 exons)	NA
ISG15	CMX- G00000000 29	CMX- V1884847	chr1:834638- 1271900	NA	CN V gain	ISG15 (2 exons)	NA
KISS1	CMX- G00000025 33	CMX- V1823995	chr1:20272910 1-205013246	NA	CN V gain	KISS1 (2 exons)	NA
KISS1R	CMX- G00000265 60	CMX- V1469394	chr19:867728- 945645	NA	CN V gain	KISS1R (2 exons)	NA
KISS1R	CMX- G00000265 60	CMX- V1974120	chr19:867728- 1126103	NA	CN V gain	KISS1R (2 exons)	NA
KISS1R	CMX- G00000265 60	CMX- V1813360	chr19:868013- 1085518	NA	CN V gain	KISS1R (2 exons)	NA
KISS1R	CMX- G00000265 60	CMX- V1883755	chr19:866589- 1232099	NA	CN V gain	KISS1R (2 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1039367	chr10:1000131 06-100022354	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1620875	chr10:1000133 59-100023161	NA	CN V loss	LOXL4 (10 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1806767	chr10:1000143 60-100020546	NA	CN V loss	LOXL4 (6 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1954806	chr10:1000141 76-100022354	NA	CN V loss	LOXL4 (8 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1107311	chr10:1000154 59-100023313	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1373344	chr10:1000154 59-100023369	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1073572	chr10:1000154 59-100023161	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4	CMX- G00000162 63	CMX- V1348325	chr10:1000145 51-100023161	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4	CMX-	CMX-	chr10:1000119	NA	CN	LOXL4 (11	NA

	G00000162 63	V1321127	10-100023369		V loss	exons)	
	CMX- G00000162 63	CMX- V1323761	chr10:1000138 76-103528663	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4							
	CMX- G00000162 63	CMX- V1275468	chr10:1000141 76-100023161	NA	CN V loss	LOXL4 (9 exons)	NA
LOXL4							
	CMX- G00000042 05	CMX- V1566424	chr2:12809360 8-128138545	NA	CN V gain	MAP3K2 (3 exons)	NA
MAP3 K2							
	CMX- G00000042 05	CMX- V1811137	chr2:12809821 6-128117112	NA	CN V gain	MAP3K2 (1 exon)	NA
MAP3 K2							
	CMX- G00000042 05	CMX- V1696049	chr2:12752027 6-128116794	NA	CN V gain	MAP3K2 (16 exons)	NA
MAP3 K2							
	CMX- G00000067 19	CMX- V9083756	chr3:19550573 9	C	T	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083757	chr3:19550596 0	G	C	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083758	chr3:19550608 9	G	A	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083759	chr3:19550609 9	T	C	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083760	chr3:19550588 3	T	C	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083761	chr3:19550114 9	C	T	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083762	chr3:19550615 6	G	C	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083763	chr3:19550589 7	G	A	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083764	chr3:19550614 6	A	G	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083765	chr3:19550614 9	C	T	Drastic nonsynonymo us	NA
MUC4							
	CMX- G00000067 19	CMX- V9083766	chr3:19550628 1	A	G	Drastic nonsynonymo us	NA
MUC4							

MUC4	CMX- G00000067 19	CMX- V9083767	chr3:19550629 1	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083768	chr3:19550630 2	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083769	chr3:19550624 5	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083770	chr3:19549591 6	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083771	chr3:19550631 8	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083772	chr3:19550632 3	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083773	chr3:19550633 9	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083774	chr3:19550635 0	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083775	chr3:19550636 4	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083776	chr3:19550618 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083777	chr3:19550619 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083778	chr3:19550639 8	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083779	chr3:19550641 0	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083780	chr3:19550641 1	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083781	chr3:19550644 6	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083782	chr3:19550646 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083783	chr3:19550600 5	A	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083784	chr3:19550652 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083785	chr3:19550653 3	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083786	chr3:19550654 2	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083787	chr3:19550578 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083788	chr3:19550655 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083789	chr3:19550659 0	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083790	chr3:19550659 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083791	chr3:19550590 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083792	chr3:19550662 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083793	chr3:19550662 7	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083794	chr3:19550674 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083795	chr3:19550674 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083796	chr3:19550649 4	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083797	chr3:19550675 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083798	chr3:19550675 2	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083799	chr3:19550675 3	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083800	chr3:19550680 9	G	T	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083801	chr3:19550691 4	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083802	chr3:19550691 7	A	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083803	chr3:19550693 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083804	chr3:19550694 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083805	chr3:19550695 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083806	chr3:19550696 5	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083807	chr3:19550696 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083808	chr3:19550697 5	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083809	chr3:19550674 7	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083810	chr3:19550698 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083811	chr3:19550698 7	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083812	chr3:19550699 0	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083813	chr3:19550701 0	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083814	chr3:19550705 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083815	chr3:19550706 2	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083816	chr3:19550637 8	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083817	chr3:19550708 3	T	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083818	chr3:19550708 6	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083819	chr3:19550710 7	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083820	chr3:19550716 6	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083821	chr3:19550720 3	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083822	chr3:19550722 6	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083823	chr3:19550722 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083824	chr3:19550723 6	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083825	chr3:19550724 2	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083826	chr3:19550725 1	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083827	chr3:19550726 2	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083828	chr3:19550731 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083829	chr3:19550732 3	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083830	chr3:19550732 4	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083831	chr3:19550736 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083832	chr3:19550737 9	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083833	chr3:19550738 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083834	chr3:19550739 7	T	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083835	chr3:19550739 8	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083836	chr3:19550740 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083837	chr3:19550741 2	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083838	chr3:19550742 2	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083839	chr3:19550742 8	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083840	chr3:19550743 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083841	chr3:19550743 4	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083842	chr3:19550744 3	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083843	chr3:19550744 5	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083844	chr3:19550744 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083845	chr3:19550746 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083846	chr3:19550747 5	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083847	chr3:19550749 1	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083848	chr3:19550749 4	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083849	chr3:19550750 2	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083850	chr3:19550760 4	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083851	chr3:19550760 5	G	A	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083852	chr3:19550761 4	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083853	chr3:19550762 0	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083854	chr3:19550762 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083855	chr3:19550763 5	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083856	chr3:19550707 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083857	chr3:19550769 4	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083858	chr3:19550773 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083859	chr3:19550777 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083860	chr3:19550779 0	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083861	chr3:19550782 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083862	chr3:19547415 9	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083863	chr3:19547778 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083864	chr3:19548900 9	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083865	chr3:19550801 9	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083866	chr3:19550802 1	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083867	chr3:19550806 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083868	chr3:19550807 0	C	T	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083869	chr3:19550809 1	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083870	chr3:19550588 6	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083871	chr3:19550811 5	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083872	chr3:19550812 7	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083873	chr3:19550590 7	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083874	chr3:19550593 0	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083875	chr3:19550595 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083876	chr3:19550833 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083877	chr3:19550597 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083878	chr3:19550845 1	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083879	chr3:19550845 3	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083880	chr3:19550847 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083881	chr3:19550847 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083882	chr3:19550850 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083883	chr3:19550850 1	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083884	chr3:19550850 2	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083885	chr3:19550852 3	C	T	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083886	chr3:19550852 6	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083887	chr3:19550866 7	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083888	chr3:19550866 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083889	chr3:19550870 2	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083890	chr3:19550631 1	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083891	chr3:19550631 5	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083892	chr3:19550878 7	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083893	chr3:19550878 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083894	chr3:19550902 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083895	chr3:19550909 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083896	chr3:19550909 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083897	chr3:19550910 2	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083898	chr3:19550638 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083899	chr3:19550921 2	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083900	chr3:19550928 7	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083901	chr3:19550935 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083902	chr3:19550935 4	C	T	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083903	chr3:19550936 3	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083904	chr3:19550936 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083905	chr3:19550937 4	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083906	chr3:19550937 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083907	chr3:19550942 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083908	chr3:19550655 4	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083909	chr3:19550956 3	A	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083910	chr3:19550957 3	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083911	chr3:19550960 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083912	chr3:19550661 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083913	chr3:19550962 7	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083914	chr3:19550965 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083915	chr3:19550975 6	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083916	chr3:19550979 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083917	chr3:19550986 1	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083918	chr3:19550987 9	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083919	chr3:19550991 8	G	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083920	chr3:19550993 9	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083921	chr3:19550994 1	A	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083922	chr3:19550995 4	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083923	chr3:19550995 7	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083924	chr3:19550997 4	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083925	chr3:19551006 8	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083926	chr3:19551008 3	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083927	chr3:19551014 6	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083928	chr3:19551019 4	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083929	chr3:19551059 0	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083930	chr3:19550698 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083931	chr3:19551065 5	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083932	chr3:19551065 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083933	chr3:19551066 2	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083934	chr3:19551068 3	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083935	chr3:19551068 6	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083936	chr3:19551069 7	G	A	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083937	chr3:19551070 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083938	chr3:19551070 7	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083939	chr3:19551070 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083940	chr3:19551071 8	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083941	chr3:19551074 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083942	chr3:19551074 9	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083943	chr3:19551076 6	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083944	chr3:19551076 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083945	chr3:19551077 3	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083946	chr3:19551082 7	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083947	chr3:19551089 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083948	chr3:19551089 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083949	chr3:19551091 0	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083950	chr3:19551094 3	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083951	chr3:19551101 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083952	chr3:19551101 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083953	chr3:19551104 3	T	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083954	chr3:19551105 1	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083955	chr3:19551107 0	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083956	chr3:19551107 6	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083957	chr3:19551110 2	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083958	chr3:19551114 2	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083959	chr3:19551115 6	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083960	chr3:19551118 6	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083961	chr3:19551119 0	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083962	chr3:19551120 4	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083963	chr3:19551121 1	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083964	chr3:19551121 4	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083965	chr3:19551126 8	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083966	chr3:19551127 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083967	chr3:19551128 5	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083968	chr3:19551128 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083969	chr3:19551133 1	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083970	chr3:19551133 6	G	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083971	chr3:19551135 8	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083972	chr3:19551139 0	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083973	chr3:19551139 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083974	chr3:19551140 3	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083975	chr3:19551141 2	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083976	chr3:19551143 8	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083977	chr3:19550768 3	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083978	chr3:19551145 4	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083979	chr3:19551146 0	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083980	chr3:19551146 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083981	chr3:19551147 4	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083982	chr3:19551148 6	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083983	chr3:19550792 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083984	chr3:19550800 9	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083985	chr3:19550801 0	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083986	chr3:19551151 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083987	chr3:19551152 5	T	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9083988	chr3:19551152 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083989	chr3:19551153 4	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083990	chr3:19551154 7	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083991	chr3:19550810 8	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083992	chr3:19551169 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083993	chr3:19551170 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083994	chr3:19550817 5	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083995	chr3:19550817 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083996	chr3:19550823 8	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083997	chr3:19551182 2	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083998	chr3:19550840 2	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9083999	chr3:19551187 0	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084000	chr3:19551187 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084001	chr3:19551191 8	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084002	chr3:19551192 5	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084003	chr3:19551193 7	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084004	chr3:19551204 2	T	C	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9084005	chr3:19551210 7	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084006	chr3:19551211 7	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084007	chr3:19551219 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084008	chr3:19551220 6	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084009	chr3:19551221 2	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084010	chr3:19551224 2	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084011	chr3:19550877 4	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084012	chr3:19550878 6	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084013	chr3:19551226 7	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084014	chr3:19551227 0	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084015	chr3:19551228 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084016	chr3:19551230 2	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084017	chr3:19551256 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084018	chr3:19551259 7	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084019	chr3:19550917 0	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084020	chr3:19551260 6	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084021	chr3:19551266 5	G	A	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9084022	chr3:19551268 6	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084023	chr3:19551269 3	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084024	chr3:19551276 7	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084025	chr3:19551276 8	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084026	chr3:19551313 6	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084027	chr3:19551315 4	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084028	chr3:19551315 5	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084029	chr3:19550947 6	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084030	chr3:19551320 3	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084031	chr3:19551321 4	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084032	chr3:19551336 4	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084033	chr3:19550961 4	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084034	chr3:19551338 3	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084035	chr3:19551339 4	A	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084036	chr3:19551339 5	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084037	chr3:19551339 7	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084038	chr3:19551339 8	C	T	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9084039	chr3:19551341 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084040	chr3:19551343 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084041	chr3:19551344 2	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084042	chr3:19551344 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084043	chr3:19551346 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084044	chr3:19551349 1	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084045	chr3:19551350 2	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084046	chr3:19551351 5	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084047	chr3:19551359 8	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084048	chr3:19551366 7	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084049	chr3:19551374 3	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084050	chr3:19551377 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084051	chr3:19551064 9	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084052	chr3:19551399 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084053	chr3:19551410 9	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084054	chr3:19551414 4	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084055	chr3:19551432 4	G	A	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9084056	chr3:19551437 9	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084057	chr3:19551440 3	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084058	chr3:19551464 3	T	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084059	chr3:19551464 5	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084060	chr3:19551464 6	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084061	chr3:19551465 4	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084062	chr3:19551466 1	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084063	chr3:19551471 8	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084064	chr3:19551472 9	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084065	chr3:19551473 3	C	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084066	chr3:19551474 1	A	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084067	chr3:19551475 7	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084068	chr3:19551480 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084069	chr3:19551481 1	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084070	chr3:19551481 2	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084071	chr3:19551482 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084072	chr3:19551484 6	A	G	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9084073	chr3:19551485 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084074	chr3:19551486 2	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084075	chr3:19551487 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084076	chr3:19551488 2	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084077	chr3:19551493 0	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084078	chr3:19551494 8	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084079	chr3:19551496 9	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084080	chr3:19551500 3	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084081	chr3:19551500 8	C	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084082	chr3:19551503 8	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084083	chr3:19551504 5	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084084	chr3:19551512 2	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084085	chr3:19551513 4	G	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084086	chr3:19551514 1	A	G	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084087	chr3:19551519 4	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084088	chr3:19551538 7	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084089	chr3:19551541 1	G	T	Drastic nonsynonymo us	NA

MUC4	CMX- G00000067 19	CMX- V9084090	chr3:19551541 3	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084091	chr3:19551544 9	A	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084092	chr3:19551545 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084093	chr3:19553890 1	C	T	Start codon gained	NA
MUC4	CMX- G00000067 19	CMX- V9084094	chr3:19551224 6	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084095	chr3:19551155 6	T	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084096	chr3:19551260 3	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084097	chr3:19551317 3	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084098	chr3:19551145 1	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084099	chr3:19551178 1	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084100	chr3:19551149 9	C	T	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084101	chr3:19551336 5	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084102	chr3:19551178 0	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084103	chr3:19551382 6	G	A	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084104	chr3:19551224 5	T	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084105	chr3:19551150 0	G	C	Drastic nonsynonymo us	NA
MUC4	CMX- G00000067 19	CMX- V9084106	chr3:19551150 2	G	C	Drastic nonsynonymo us	NA

MUC4	CMX-G0000006719	CMX-V9084107	chr3:195511859	T	G	Drastic nonsynonymous	NA
MUC4	CMX-G0000006719	CMX-V9084108	chr3:195511783	A	G	Drastic nonsynonymous	NA
MUC4	CMX-G0000006719	CMX-SV00002	chr3:195512373	G	GGAT	Codon change and codon insertion	NA
MUC4	CMX-G0000006719	CMX-SV00003	chr3:195518112	T	TGTCTCTGCGTAA CA	Codon change and codon insertion	NA
MUC4	CMX-G0000006719	CMX-SV00004	chr3:195464985	CNV duplication	NA	Splice acceptor variant	NA
MUC4	CMX-G0000006719	CMX-SV00005	chr3:195507809	CNV deletion	NA	Nonsynonymous and coding sequence	NA
MUC4	CMX-G0000006719	CMX-SV00006	chr3:195508499	CNV duplication	NA	Frameshift	NA
MUC4	CMX-G0000006719	CMX-V9084187	chr3:195499847	A	G	NA	6.75E-05
MUC4	CMX-G0000006719	CMX-V9084188	chr3:195500367	A	G	NA	0.000532509
MUC4	CMX-G0000006719	CMX-V9084191	chr3:195506750	G	C	NA	0.000425548
MUC4	CMX-G0000006719	CMX-V9084192	chr3:195506760	T	A	NA	7.68E-05
MUC4	CMX-G0000006719	CMX-V9084189	chr3:195506195	C	T	NA	8.00E-05
MUC4	CMX-G0000006719	CMX-V9084190	chr3:195506746	G	A	NA	0.000150373
NLRP11	CMX-G00000028188	CMX-V9084110	chr19:56320663	G	A	Drastic nonsynonymous	NA
NLRP11	CMX-G00000028188	CMX-V9084111	chr19:56329447	G	A	Drastic nonsynonymous	NA
NLRP11	CMX-G00000028188	CMX-V9084112	chr19:56343378	C	A	Start codon gained	NA

NLRP1 4	CMX- G00000169 19	CMX- V9084115	chr11:7091569	C	T	Drastic nonsynonymo us	NA
NLRP1 4	CMX- G00000169 19	CMX- V9084116	chr11:7079038	G	A	Drastic nonsynonymo us	NA
NLRP1 4	CMX- G00000169 19	CMX- V9084117	chr11:7059981	G	A	Drastic nonsynonymo us	NA
NLRP5	CMX- G00000281 92	CMX- V9084120	chr19:5656962 9	C	G	Drastic nonsynonymo us	NA
NLRP5	CMX- G00000281 92	CMX- V9084121	chr19:5657287 5	G	A	Drastic nonsynonymo us	NA
NLRP5	CMX- G00000281 92	CMX- V9084170	chr19:5656714 7	A	G	NA	8.96E- 06
NLRP5	CMX- G00000281 92	CMX- V9084169	chr19:5656713 3	A	G	NA	0.0004 22755
NLRP8	CMX- G00000281 91	CMX- V9084122	chr19:5645934 2	C	T	Drastic nonsynonymo us	NA
NLRP8	CMX- G00000281 91	CMX- V9084123	chr19:5646737 5	C	T	Drastic nonsynonymo us	NA
NLRP8	CMX- G00000281 91	CMX- V9084124	chr19:5649927 9	G	C	Stop codon lost	NA
PADI3	CMX- G00000003 42	CMX- V1792728	chr1:17548826 -18037716	NA	CN V gain	PADI3 (16 exons)	NA
PADI6	CMX- G00000003 44	CMX- V9084147	chr1:17707931	T	G	NA	0.0009 47202
PADI6	CMX- G00000003 44	CMX- V9084145	chr1:17707757	C	T	NA	0.0007 91492
PADI6	CMX- G00000003 44	CMX- V9084146	chr1:17707758	G	C	NA	0.0008 32422
PAEP	CMX- G00000152 54	CMX- V1271620	chr9:13813147 6-138644038	NA	CN V gain	PAEP (2 exons)	NA
PLCB1	CMX- G00000284 45	CMX- V1930635	chr20:8142398 -10362561	NA	CN V gain	PLCB1 (2 exons)	NA
PMS2	CMX- G00000112 51	CMX- V9084128	chr7:6045627	C	T	Drastic nonsynonymo us	NA

PMS2	CMX-G0000011251	CMX-SV00007	chr7:6029313	CNV duplication	NA	Splice donor, acceptor and coding sequence	NA
PMS2	CMX-G0000011251	CMX-V9084222	chr7:5981433	A	G	NA	0.000681822
POF1B	CMX-G0000031099	CMX-V1507096	chrX:77243971-85734966	NA	CNV gain	POF1B (15 exons)	NA
PRDM9	CMX-G0000008219	CMX-V1222200	chr5:21969693-23940832	NA	CNV loss	PRDM9 (3 exons)	NA
SCARB1	CMX-G0000019991	CMX-V9084131	chr12:125270773	A	G	Drastic nonsynonymous	NA
SCARB1	CMX-G0000019991	CMX-V9084132	chr12:125323962	A	C	Start codon gained	NA
SCARB1	CMX-G0000019991	CMX-V9084133	chr12:125324570	C	T	Start codon gained	NA
SCARB1	CMX-G0000019991	CMX-V9084134	chr12:125324553	C	T	Start codon gained	NA
SERPINA10	CMX-G0000021629	CMX-V1143735	chr14:94691918-95251285	NA	CNV gain	SERPINA10 (4 exons)	NA
SIRT3	CMX-G0000016629	CMX-V1733950	chr11:222921-278027	NA	CNV loss	SIRT3 (2 exons)	NA
SPIN1	CMX-G0000014689	CMX-V9084227	chr9:90754700	G	A	NA	0.000183378
SPIN1	CMX-G0000014689	CMX-V9084228	chr9:90754733	A	C	NA	0.000548473
SPIN1	CMX-G0000014689	CMX-V9084229	chr9:91120108	G	A	NA	0.000742923
SPIN1	CMX-G0000014689	CMX-V9084230	chr9:91120393	A	G	NA	0.000742923
SPIN1	CMX-G0000014689	CMX-V9084231	chr9:91124743	A	G	NA	0.000742923
SPIN1	CMX-G0000014689	CMX-V9084232	chr9:91126304	C	T	NA	0.000742923
SPIN1	CMX-G0000014689	CMX-V9084233	chr9:91126736	G	A	NA	0.00031089

	89						
SPIN1	CMX- G00000146 89	CMX- V9084234	chr9:91130846	G	A	NA	0.0007 71149
SPIN1	CMX- G00000146 89	CMX- V9084235	chr9:91131392	A	G	NA	0.0009 34759
SPIN1	CMX- G00000146 89	CMX- V9084236	chr9:91133854	T	A	NA	0.0008 58194
SPIN1	CMX- G00000146 89	CMX- V9084237	chr9:91139780	C	T	NA	0.0009 10019
SPIN1	CMX- G00000146 89	CMX- V9084238	chr9:91146391	C	T	NA	0.0004 84881
SPN	CMX- G00000236 64	CMX- V1697382	chr16:2927495 5-29761984	NA	CN V loss	SPN (1 exon)	NA
TACC3	CMX- G00000068 18	CMX- V9084137	chr4:1729556	G	A	Drastic nonsynonymo us	NA
TACC3	CMX- G00000068 18	CMX- V9084138	chr4:1732978	G	A	Drastic nonsynonymo us	NA
TLE6	CMX- G00000266 39	CMX- V1806717	chr19:2946999 -3051118	NA	CN V loss	TLE6 (2 exons)	NA
TLE6	CMX- G00000266 39	CMX- V1336365	chr19:2937389 -3057790	NA	CN V loss	TLE6 (2 exons)	NA
ZP3	CMX- G00000119 47	CMX- V9084143	chr7:76058767	G	T	Start codon gained	NA
NA	NA	CMX- V2992389	chr1:3584692- 3585200	NA	CN V gain	NA	0.0003 63085
NA	NA	CMX- V2992390	chr1:33214881 -33216355	NA	CN V loss	NA	0.0014 5087
NA	NA	CMX- V2992391	chr1:11025279 2-110252792	NA	CN V loss	NA	0.0014 5087
NA	NA	CMX- V2992392	chr1:14880005 6-148802742	NA	CN V gain	NA	0.0003 63942
NA	NA	CMX- V2992393	chr2:86414923 -86421116	NA	CN V loss	NA	0.0014 5087
NA	NA	CMX-	chr2:96237124	NA	CN	NA	1.3320

		V2992394	-96237180		V		7E-05
					gain		
					CN		
NA	NA	CMX-	chr2:21540426		V		0.0002
		V2992395	0-215412550	NA	loss	NA	69506
					CN		
NA	NA	CMX-	chr2:21721072		V		0.0014
		V2992396	0-217210773	NA	loss	NA	1334
					CN		
NA	NA	CMX-	chr3:38475943		V		0.0002
		V2992397	-38476013	NA	loss	NA	63066
					CN		
NA	NA	CMX-	chr3:15057714		V		0.0014
		V2992398	8-150583696	NA	loss	NA	5087
					CN		
NA	NA	CMX-	chr4:95892431		V		0.0005
		V2992399	-95892748	NA	loss	NA	95928
					CN		
NA	NA	CMX-	chr4:10396529		V		9.3208
		V2992400	6-103966620	NA	gain	NA	4E-05
					CN		
NA	NA	CMX-	chr4:17469163		V		0.0010
		V2992401	3-174691747	NA	loss	NA	24494
					CN		
NA	NA	CMX-	chr5:10634995		V		0.0016
		V2992402	0-106350159	NA	loss	NA	66446
					CN		
NA	NA	CMX-	chr5:17965488		V		0.0009
		V2992403	3-179655477	NA	loss	NA	1471
					CN		
NA	NA	CMX-	chr6:77073676		V		0.0001
		V2992404	-77085224	NA	gain	NA	0917
					CN		
NA	NA	CMX-	chr7:43968000		V		0.0008
		V2992405	-44039304	NA	loss	NA	60892
					CN		
NA	NA	CMX-	chr7:69794356		V		0.0014
		V2992406	-69800088	NA	loss	NA	5087
					CN		
NA	NA	CMX-	chr7:99464961		V		0.0012
		V2992407	-99465782	NA	loss	NA	5626
					CN		
NA	NA	CMX-	chr7:10171397		V		0.0008
		V2992408	7-101923980	NA	loss	NA	60892
					CN		
NA	NA	CMX-	chr8:12292467		V		0.0011
		V2992409	-12292467	NA	gain	NA	6959
					CN		
NA	NA	CMX-	chr8:14172343		V		0.0014
		V2992410	6-141723436	NA	loss	NA	19478

NA	NA	CMX-V2992411	chr8:14546500	NA	CN V loss	NA	0.0004 88267
NA	NA	CMX-V2992412	chr9:11921363	NA	NA	NA	0.0014 46882
NA	NA	CMX-V2992413	chr9:12919995	NA	CN V gain	NA	0.0004 6153
NA	NA	CMX-V2992414	chr9:13855781	NA	CN V loss	NA	0.0014 46882
NA	NA	CMX-V2992415	chr10:1342520	NA	CN V loss	NA	0.0002 95719
NA	NA	CMX-V2992416	chr10:7935275	NA	CN V loss	NA	0.0014 5087
NA	NA	CMX-V2992417	chr10:1350379	NA	CN V loss	NA	0.0009 83276
NA	NA	CMX-V2992418	chr11:2113479	NA	CN V loss	NA	0.0015 66125
NA	NA	CMX-V2992419	chr11:2052165	NA	CN V loss	NA	0.0014 45217
NA	NA	CMX-V2992420	chr11:7216534	NA	CN V loss	NA	0.0003 66026
NA	NA	CMX-V2992421	chr12:1103363	NA	CN V loss	NA	0.0002 63066
NA	NA	CMX-V2992422	chr12:1315801	NA	CN V loss	NA	0.0004 34354
NA	NA	CMX-V2992423	chr13:1059829	NA	CN V loss	NA	0.0015 66125
NA	NA	CMX-V2992424	chr14:1047118	NA	CN V loss	NA	0.0001 17224
NA	NA	CMX-V2992425	chr14:1055548	NA	CN V gain	NA	0.0011 5304
NA	NA	CMX-V2992426	chr14:1060381	NA	CN V gain	NA	0.0013 88783
NA	NA	CMX-V2992427	chr15:7247390	NA	CN V gain	NA	2.2682 E-05

NA	NA	CMX-V2992428	chr15:8174301-1-81748883	NA	CN V loss	NA	0.0009 34763
NA	NA	CMX-V2992429	chr15:9700621-1-97006211	NA	CN V loss	NA	0.0008 8514
NA	NA	CMX-V2992430	chr16:420035-420035	NA	CN V loss	NA	0.0010 33484
NA	NA	CMX-V2992431	chr16:2829796-2-28340178	NA	CN V loss	NA	3.8376 9E-05
NA	NA	CMX-V2992432	chr16:2861400-7-28653740	NA	CN V loss	NA	0.0003 37601
NA	NA	CMX-V2992433	chr16:3377293-6-33809650	NA	CN V loss	NA	0.0012 24595
NA	NA	CMX-V2992434	chr17:3768689-2-37687211	NA	CN V loss	NA	0.0002 63066
NA	NA	CMX-V2992435	chr17:7036567-3-70365673	NA	CN V loss	NA	0.0016 52185
NA	NA	CMX-V2992436	chr17:7741878-9-77465794	NA	CN V loss	NA	0.0001 17224
NA	NA	CMX-V2992437	chr19:1532671-1549096	NA	CN V loss	NA	0.0009 34076
NA	NA	CMX-V2992438	chr19:1883556-2-18835562	NA	CN V loss	NA	0.0012 24595
NA	NA	CMX-V2992439	chr19:3848019-9-38480199	NA	CN V loss	NA	0.0002 69506
NA	NA	CMX-V2992440	chr19:4573178-5-45732555	NA	CN V loss	NA	0.0002 29579
NA	NA	CMX-V2992441	chr19:5310200-0-53153808	NA	CN V gain	NA	0.0016 44428
NA	NA	CMX-V2992442	chr20:1500411-1508282	NA	CN V loss	NA	0.0004 61106
NA	NA	CMX-V2992443	chr20:6694925-6696738	NA	CN V loss	NA	0.0009 34763
NA	NA	CMX-V2992444	chr20:6159220-2-61594834	NA	CN V loss	NA	0.0014 94022

NA	NA	CMX-V2992445	chr21:15355967-15355967	NA	CN V loss	NA	0.001566125
NA	NA	CMX-V2992446	chr21:44541166-44547084	NA	CN V loss	NA	0.000257622
NA	NA	CMX-V2992447	chrX:100110102-100110152	NA	CN V loss	NA	0.001445217
NA	NA	CMX-V2992448	chrX:152934795-152944222	NA	CN V loss	NA	0.000247877

Description of Certain Genes

- 5 Below are detailed descriptions of some of the fertility genes described in the tables above.

BARD1

- BRCA1-Associated Ring Domain 1 (BARD1) is a gene that forms a heterodimer complex with the BRCA1 gene, and this complex is required for spindle-pole assembly in mitosis, and hence
 10 chromosome stability. Mouse embryos carrying homozygous null alleles for BARD1 died between embryonic day 7.5 and embryonic day 8.5 due to severely impaired cell proliferation (McCarthy et al. Molec. Cell. Biol. 23: 5056-5063, 2003).

C6orf221 (KHDC3L)

- 15 KH domain containing 3-like, subcortical maternal complex member (KHDC3L). The gene also has the identifier "C6orf221" [Entrez Gene id: 154288 , HGNC id: 33699]. KH domains are protein domains that binds to RNA molecules, and KHDC3L is likely involved in genomic imprinting, a phenomenon where genes are expressed in a parental-origin specific manner. KHDC3L gene expression is maximal in germinal vesicle oocytes, tailing off through metaphase
 20 II oocytes, and its expression profile is similar to other oocyte-specific genes [Am J Hum Genet. 2011 September 9; 89(3): 451–458]. It is also found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23]. Mice carrying homozygous null alleles for KHDC3L display a maternal effect defect in embryogenesis with delayed embryonic development and spindle abnormalities
 25 resulting in decreased litter sizes for homozygous females. In humans, KHDC3L has been

implicated in familial biparental hydatidiform mole, a maternal-effect recessive inherited disorder [Ref: Am J Hum Genet. 2011 September 9; 89(3): 451–458]

DNMT1

5 DNA (cytosine-5)-methyltransferase 1 (DNMT1) [Entrez Gene id: 1786, HGNC id: 2976], belongs to a group of enzymes that transfer methyl groups to position 5 of cytosine bases in DNA. While this process, known as DNA methylation, does not alter DNA base composition, it leaves “epigenetic” modifications to DNA molecules that affect the biochemical properties of the DNA region. DNA methylation, mediated by DNMT1, is crucial in determining cell fate during
10 embryogenesis [Genes Dev. 2008 Jun 15;22(12):1607-16, Dev Biol. 2002 Jan 1;241(1):172-82.]. Mouse embryos carrying homozygous null alleles for DNMT1 survive only to mid-gestation. The expression of the DNMT1 gene is significantly higher in reproductive tissues than other cell types, and is found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23, BMC Genomics.
15 2009 Aug 3;10:348].

FMR1

Fragile X Mental Retardation 1 (FMR1) encodes for the RNA-binding protein FMRP that is implicated in the fragile-X syndrome. The inhibition of translation may be a function of FMR1
20 in vivo, and that failure of mutant FMR1 protein to oligomerize may contribute to the pathophysiologic events leading to fragile X syndrome. Fragile X premutations in female carriers appear to be a risk factor for premature ovarian failure: 16% of the premutation carriers, menopause occurred before the age of 40, compared with none of the full-mutation carriers and 1 (0.4%) of the controls, indicating a significant association between premature menopause and
25 premutation carrier status. [Am. J. Med. Genet. 83: 322-325, 1999]

FOXO3

Foxhead box O3 (FOXO3) encodes a protein that induces apoptosis in cells, lying within the DNA damage response and repair pathways. FOXO3 knockout female mice exhibit infertility
30 phenotypes, in particular abnormal ovarian follicular function. Mice mutants carrying a

homozygous non-synonymous substitution in exon 2 of the FOXO3 gene show loss of fertility of sexual maturity and exhibit premature ovarian failures. [Mammalian Genome 22: 235-248, 2011]

MUC4

- 5 MUC4 belongs to a family of high-molecular-weight glycoproteins that protect and lubricate the epithelial surface of respiratory, gastrointestinal and reproductive tracts. The extracellular domain can interact with an epidermal growth factor receptor on the cell surface to modulate downstream cell growth signaling by stabilizing and/or enhancing the activity of cell growth receptor complexes [Nature Rev. Cancer. 4(1):45-60, 2004]. MUC4 is expressed in the
- 10 endometrial epithelium and is associated with endometriosis development and endometriosis-related infertility such as embryo implantation [BMC Med. 2011 9:19, 2011].

NLRP11

- NLR family, pyrin domain containing 11 (NLRP11) encodes a leucine-rich protein belonging to
- 15 a large family of proteins likely involved in inflammation [Nature Rev. Molec. Cell Biol. 4: 95-104, 2003], and is expressed in the ovary, testes and pre-implantation embryos [BMC Evol Biol. 2009 Aug 14;9:202. doi: 10.1186/1471-2148-9-202.]. NLRP11 gene expression shows specificity to reproductive tissues.

NLRP14

- NLR family, pyrin domain containing 14 (NLRP14) encodes a leucine-rich protein belonging to a large family of proteins likely involved in inflammation [Nature Rev. Molec. Cell Biol. 4: 95-104, 2003], and is expressed in the ovary, testes and pre-implantation embryos [BMC Evol Biol. 2009 Aug 14;9:202. doi: 10.1186/1471-2148-9-202.]. NPRL14 is also found within the set of
- 25 maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23, BMC Genomics. 2009 Aug 3;10:348].

NLRP5

- NLRP5 or MATER (Maternal antigen the embryos require), the protein
- 30 encoded by the *Nlrp5* gene, is another highly abundant oocyte protein that is essential in mouse for embryonic development beyond the two-cell stage. MATER was originally identified as an

oocyte-specific antigen in a mouse model of autoimmune premature ovarian failure (Tong et al., 25 Endocrinology, 140:3720-3726, 1999). MATER demonstrates a similar expression and subcellular expression profile to PADI6. Like *Padi6*-null animals, *Nlrp5*-null females exhibit normal oogenesis, ovarian development, oocyte maturation, ovulation and fertilization. However, 5 embryos derived from *Nlrp5*-null females undergo a developmental block at the two-cell stage and fail to exhibit normal embryonic genome activation (Tong et al., Nat Genet 26:267-268, 2000; and Tong et al. Mamm Genome 11:281-287, 2000b).

NLRP8

10 NLR family, pyrin domain containing 8 (NLRP8) encodes a leucine-rich protein belonging to a large family of proteins likely involved in inflammation [Nature Rev. Molec. Cell Biol. 4: 95-104, 2003], and is expressed in the ovary, testes and pre-implantation embryos [BMC Evol Biol. 2009 Aug 14;9:202. doi: 10.1186/1471-2148-9-202.]. NLRP8 gene expression shows specificity to reproductive tissues.

NPM2

The gene NPM2 [Entrez Gene id : 10361, HGNC id: 7930], or nucleoplasmin 2, is a chaperon that binds to histones, and is involved in sperm chromatin remodeling after oocyte entry [Nucleic Acids Res. 2012 June; 40(11): 4861–4878]. NPM2 has been found in a screen for oocyte-specific 20 genes involved in preimplantation embryonic development [Semin Reprod Med. 2007 Jul;25(4):243-51], and is differentially expressed during final oocyte maturation and early embryonic development in humans [Fertil Steril. 2007 Mar;87(3):677-90]. NPM2 is a maternal effect gene critical for nuclear and nucleolar organization and embryonic development, and is found within the set of maternal factors that are important for driving egg-to-embryo transition 25 during fertilization [Reproduction. 2010 May;139(5):809-23, BMC Genomics. 2009 Aug 3;10:348]. NPM2 is associated with abnormal oocyte morphology and reduced fertility in mice, and female mice homozygous null for NPM2 carry defects in preimplantation embryo development, with abnormalities in oocyte and early embryonic nuclei [Science. 2003 Apr 25;300(5619):633-6].

PADI6

Peptidylarginine deiminase 6 (*PADI6*) *Padi6* was originally cloned from a 2D murine egg proteome gel based on its relative abundance, and *Padi6* expression in mice appears to be almost entirely limited to the oocyte and pre-implantation embryo (Yurttas et al., 2010). *Padi6* is first expressed in primordial oocyte follicles and persists, at the protein level, throughout pre-implantation development to the blastocyst stage (Wright et al., Dev Biol, 256:73-88, 2003). Inactivation of *Padi6* leads to female infertility in mice, with the *Padi6*-null developmental arrest occurring at the two-cell stage (Yurttas et al., 2008).

PMS2

10 PMS2 is involved in DNA mismatch repair and involved in fertilization and pre-implantation development. It has been identified by knockout mouse studies as one of many maternal effect genes essential for development [Nature Cell Bio. 4 Suppl, pp.s41–9].

SCARB1

15 Scavenger receptor class B, member 1 (*SCARB1*) gene encodes a glycoprotein that is a receptor for mediating cholesterol transport. *SCARB1*-null homozygous female mice were infertile with dysfunctional oocytes [J. Clin. Invest. 108: 1717-1722, 2001], hence, mutations in *SCARB1* may affect female fertility by regulating lipoprotein metabolism.

SPIN1

20 Spindlin 1 (*SPIN1*) is a gene abundantly expressed in early embryo development, during the transition from oocyte to pluripotent early-embryo. *SPIN1* is phosphorylated in a cell-cycle dependent manner and is associated with the meiotic spindle [Development 124: 493-503, 1997].

TACC3

25 Transforming, Acidic Coiled-Coil Containing Protein 3 (*TACC3*). In mice, *TACC3* is abundantly expressed in the cytoplasm of growing oocytes, and is required for microtubule anchoring at the centrosome and for spindle assembly and cell survival (Fu et al., 2010). *TACC3* is also found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23, BMC Genomics. 2009 Aug 3;10:348].

30

ZP1

Zona pellucid glycoprotein 1 (ZP1) encodes for a protein that is a structural component of the zona pellucida - an extracellular matrix that surrounds the oocyte and early embryo.

5 **ZP2**

Zona pellucid glycoprotein 2 (ZP2) encodes for a protein that is a structural component of the zona pellucida - an extracellular matrix that surrounds the oocyte and early embryo. ZP2 binds to acrosome-reacted sperm and is important in preventing polyspermy [Hum Reprod. 2004 Jul;19(7):1580-6.].

10

ZP3

Zona pellucid glycoprotein 3 (ZP3) [Entrez Gene id : 7784, HGNC id: 13189], is a structural component of the zona pellucida - an extracellular matrix that surrounds the oocyte and early embryo. It is found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [BMC Genomics. 2009 Aug 3;10:348]. ZP3 is also expressed in oocytes from early ovarian development, and likely to have a role in the development of primordial follicle before zona pellucida formation [Mol Cell Endocrinol. 2008 Jul 16;289(1-2):10-5]. Female mice carrying null alleles for ZP3 exhibit decreased ovary size and weight, abnormal ovarian folliculogenesis and ovulation, ultimately resulting in female infertility.

20

ZP4

Zona pellucid glycoprotein 4 (ZP4) encodes for a protein that is a structural component of the zona pellucida - an extracellular matrix that surrounds the oocyte and early embryo. ZP4 stimulates acrosome reaction as part of a signaling pathway that involves Protein Kinase A [Biol Reprod. 2008 Nov;79(5):869-77]

25

DNA (cytosine-5)-methyltransferase 1 (DNMT1) [Entrez Gene id: 1786, HGNC id: 2976], belongs to a group of enzymes that transfer methyl groups to position 5 of cytosine bases in DNA. While this process, known as DNA methylation, does not alter DNA base composition, it leaves “epigenetic” modifications to DNA molecules that affect the biochemical properties of the

30

DNA region. DNA methylation, mediated by DNMT1, is crucial in determining cell fate during embryogenesis [Genes Dev. 2008 Jun 15;22(12):1607-16, Dev Biol. 2002 Jan 1;241(1):172-82.]. Mouse embryos carrying homozygous null alleles for DNMT1 survive only to mid-gestation. The expression of the DNMT1 gene is significantly higher in reproductive tissues than other cell types, and is found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23, BMC Genomics. 2009 Aug 3;10:348].

The gene **NPM2** [Entrez Gene id : 10361, HGNC id: 7930], or nucleoplasmin 2, is a chaperon that binds to histones, and is involved in sperm chromatin remodeling after oocyte entry [Nucleic Acids Res. 2012 June; 40(11): 4861–4878]. NPM2 has been found in a screen for oocyte-specific genes involved in preimplantation embryonic development [Semin Reprod Med. 2007 Jul;25(4):243-51], and is differentially expressed during final oocyte maturation and early embryonic development in humans [Fertil Steril. 2007 Mar;87(3):677-90]. NPM2 is a maternal effect gene critical for nuclear and nucleolar organization and embryonic development, and is found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23, BMC Genomics. 2009 Aug 3;10:348]. NPM2 is associated with abnormal oocyte morphology and reduced fertility in mice, and female mice homozygous null for NPM2 carry defects in preimplantation embryo development, with abnormalities in oocyte and early embryonic nuclei [Science. 2003 Apr 25;300(5619):633-6].

Oocyte-Expressed Protein (OOEP) [Entrez Gene id : 441161, HGNC id: 21382], also goes by the identifiers KHDC2, FLOPED, HOEP19 and C6orf156. OOEP is found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [Reproduction. 2010 May;139(5):809-23]. OOEP is expressed in ovaries, but not detectable in other cell types including male testes. Within the ovary, its expression is restricted to growing oocytes. The OOEP protein product sublocalizes to the subcortex of eggs and preimplantation embryos. OOEP homozygous null female mice have seemingly normal ovarian physiology and produced viable eggs that can be fertilized, however, these embryos do not progress beyond cleavage stage development and hence these female mice are sterile. It is believed that a functioning OOEP is a pre-requisite for pre-implantation mouse development [Dev Cell. 2008 September; 15(3): 416–425.].

Factor located in oocytes permitting embryonic development (FLOPED /OOEP) The subcortical maternal complex (SCMC) is a poorly characterized murine oocyte structure to which several maternal effect gene products localize (Li et al. Dev Cell 15:416-425, 2008). PADI6, MATER, FILIA, TLE6, and FLOPED have been shown to localize to this complex (Li et al. Dev Cell 15:416-425, 2008; Yurttas et al. Development 135:2627-2636, 2008). This complex is not present in the absence of *Floped* and *Nlrp5*, and similar to embryos resulting from *Nlrp5*-depleted oocytes, embryos resulting from *Floped*-null oocytes do not progress past the two cell stage of mouse development (Li et al., 2008). FLOPED is a small (19kD) RNA binding protein that has also been characterized under the name of MOEP19 (Herr et al., Dev Biol 314:300-316, 2008).

Zona pellucid glycoprotein 3 (ZP3) [Entrez Gene id : 7784, HGNC id: 13189], is a structural component of the zona pellucida - an extracellular matrix that surrounds the oocyte and early embryo. It is found within the set of maternal factors that are important for driving egg-to-embryo transition during fertilization [BMC Genomics. 2009 Aug 3;10:348]. ZP3 is also expressed in oocytes from early ovarian development, and likely to have a role in the development of primordial follicle before zona pellucida formation [Mol Cell Endocrinol. 2008 Jul 16;289(1-2):10-5]. Female mice carrying null alleles for ZP3 exhibit decreased ovary size and weight, abnormal ovarian folliculogenesis and ovulation, ultimately resulting in female infertility.

FIGLA (Factor in Germline Alpha) [Entrez Gene id : 344018 , HGNC id:], also goes by the gene identifiers POF6, BHLHC8, and FIGALPHA. This gene is a basic helix-loop-helix transcription factor that acts as an activator of oocyte genes. FIGLA is expressed in all ovarian follicular stages and in mature oocytes, and is required for normal folliculogenesis. FIGLA expression is also believed to repress genes expressed normal in male testes, and hence sustains the female phenotype by activating female and repressing male germ cell genetic hierarchies in growing oocytes during postnatal ovarian development [Mol Cell Biol. 2010 July; 30(14)]. Female mice with FIGLA mutations result in decreased oocytes numbers and abnormal ovarian folliculogenesis. Heterozygous mutations in FIGLA has been implicated in women with premature ovarian failure [Am J Hum Genet. 2008 Jun;82(6):1342-8.].

Peptidylarginine deiminase 6 (PADI6) *Padi6* was originally cloned from a 2D murine egg proteome gel based on its relative abundance, and *Padi6* expression in mice appears to be

almost entirely limited to the oocyte and pre-implantation embryo (Yurttas et al., 2010). *Padi6* is first expressed in primordial oocyte follicles and persists, at the protein level, throughout pre-implantation development to the blastocyst stage (Wright et al., Dev Biol, 256:73-88, 2003). Inactivation of *Padi6* leads to female infertility in mice, with the *Padi6*-null developmental arrest occurring at the two-cell stage (Yurttas et al., 2008).

Maternal antigen the embryos require (MATER / *NLRP5*) MATER, the protein encoded by the *Nlrp5* gene, is another highly abundant oocyte protein that is essential in mouse for embryonic development beyond the two-cell stage. MATER was originally identified as an oocyte-specific antigen in a mouse model of autoimmune premature ovarian failure (Tong et al., Endocrinology, 140:3720-3726, 1999). MATER demonstrates a similar expression and subcellular expression profile to PADI6. Like *Padi6*-null animals, *Nlrp5*-null females exhibit normal oogenesis, ovarian development, oocyte maturation, ovulation and fertilization. However, embryos derived from *Nlrp5*-null females undergo a developmental block at the two-cell stage and fail to exhibit normal embryonic genome activation (Tong et al., Nat Genet 26:267-268, 2000; and Tong et al. Mamm Genome 11:281-287, 2000b).

KH domain containing 3-like, subcortical maternal complex member (FILIA/*KHDC3L*) FILIA is another small RNA-binding domain containing maternally inherited murine protein. FILIA was identified and named for its interaction with MATER (Ohsugi et al. Development 135:259-269, 2008). Like other components of the SCMC, maternal inheritance of the *Khdc3* gene product is required for early embryonic development. In mice, loss of *Khdc3* results in a developmental arrest of varying severity with a high incidence of aneuploidy due, in part, to improper chromosome alignment during early cleavage divisions (Li et al., 2008). *Khdc3* depletion also results in aneuploidy, due to spindle checkpoint assembly (SAC) inactivation, abnormal spindle assembly, and chromosome misalignment (Zheng et al. Proc Natl Acad Sci USA 106:7473-7478, 2009).

Basonuclin (*BNCT*) Basonuclin is a zinc finger transcription factor that has been studied in mice. It is found expressed in keratinocytes and germ cells (male and female) and regulates rRNA (via polymerase I) and mRNA (via polymerase II) synthesis (Iuchi and Green, 1999; Wang et al., 2006). Depending on the amount by which expression is reduced in oocytes, embryos may not develop beyond the 8-cell stage. In *Bsn1* depleted mice, a normal number of

oocytes are ovulated even though oocyte development is perturbed, but many of these oocytes cannot go on to yield viable offspring (Ma et al., 2006).

Zygote Arrest 1 (ZAR1) *Zar1* is an oocyte-specific maternal effect gene that is known to function at the oocyte to embryo transition in mice. High levels of *Zar1* expression are observed in the cytoplasm of murine oocytes, and homozygous-null females are infertile: growing oocytes from *Zar1*-null females do not progress past the two-cell stage.

Cytosolic phospholipase A2 γ (PLA2G4C) Under normal conditions, cPLA2 γ , the protein product of the murine *PLA2G4C* ortholog, expression is restricted to oocytes and early embryos in mice. At the subcellular level, cPLA2 γ mainly localizes to the cortical regions, nucleoplasm, and multivesicular aggregates of oocytes. It is also worth noting that while cPLA2 γ expression does appear to be mainly limited to oocytes and pre-implantation embryos in healthy mice, expression is considerably up-regulated within the intestinal epithelium of mice infected with *Trichinella spiralis*. This suggests that cPLA2 γ may also play a role in the inflammatory response. The human PLA2G4C differs in that rather than being abundantly expressed in the ovary, it is abundantly expressed in the heart and skeletal muscle. Also, the human protein contains a lipase consensus sequence but lacks a calcium-binding domain found in other PLA2 enzymes. Accordingly, another cytosolic phospholipase may be more relevant for human fertility.

Transforming, Acidic Coiled-Coil Containing Protein 3 (TACC3) In mice, TACC3 is abundantly expressed in the cytoplasm of growing oocytes, and is required for microtubule anchoring at the centrosome and for spindle assembly and cell survival (Fu et al., 2010). In certain embodiments, the gene is a gene that is expressed in an oocyte. Exemplary genes include *CTCF*, *ZFP57*, *POU5F1*, *SEBOX*, and *HDAC1*.

In other embodiments, the gene is a gene that is involved in DNA repair pathways, including but not limited to, *MLH1*, *PMS1* and *PMS2*. In other embodiments, the gene is *BRCA1* or *BRCA2*.

In other embodiments, the biomarker is a gene product (e.g., RNA or protein) of an infertility-associated gene. In particular embodiments, the gene product is a gene product of a maternal effect gene. In other embodiments, the gene product is a product of a gene from Table 1. In certain embodiments, the gene product is a product of a gene that is expressed in an oocyte, such as a product of *CTCF*, *ZFP57*, *POU5F1*, *SEBOX*, and *HDAC1*. In other embodiments, the

gene product is a product of a gene that is involved in DNA repair pathways, such as a product of *MLH1*, *PMS1*, or *PMS2*. In other embodiments, gene product is a product of *BRCA1* or *BRCA2*.

In other embodiments, the biomarker may be an epigenetic factor, such as methylation patterns (e.g., hypermethylation of CpG islands), genomic localization or post-translational
5 modification of histone proteins, or general post-translational modification of proteins such as acetylation, ubiquitination, phosphorylation, or others.

In other embodiments, methods of the invention analyze infertility-associated biomarkers in order to assess the risk infertility.

In certain embodiments, the biomarker is a genetic region, gene, or RNA/protein product
10 of a gene associated with the one carbon metabolism pathway and other pathways that effect methylation of cellular macromolecules. Exemplary genes and products of those genes are described below.

Methylenetetrahydrofolate Reductase (*MTHFR*) In particular embodiments a mutation (677C>T) in the *MTHFR* gene is associated with infertility. The enzyme 5,10-
15 methylenetetrahydrofolate reductase regulates folate activity (Pavlik et al., Fertility and Sterility 95(7): 2257-2262, 2011). The 677TT genotype is known in the art to be associated with 60% reduced enzyme activity, inefficient folate metabolism, decreased blood folate, elevated plasma homocysteine levels, and reduced methylation capacity. Pavlik et al. (2011) investigated the effect of the *MTHFR* 677C>T on serum anti-Mullerian hormone (AMH) concentrations and on
20 the numbers of oocytes retrieved (NOR) following controlled ovarian hyperstimulation (COH). Two hundred and seventy women undergoing COH for IVF were analyzed, and their AMH levels were determined from blood samples collected after 10 days of GnRH superagonist treatment and before COH. Average AMH levels of TT carriers were significantly higher than those of homozygous CC or heterozygous CT individuals. AMH serum concentrations correlated
25 significantly with the NOR in all individuals studied. The study concluded that the *MTHFR* 677TT genotype is associated with higher serum AMH concentrations but paradoxically has a negative effect on NOR after COH. It was proposed that follicle maturation might be retarded in *MTHFR* 677TT individuals, which could subsequently lead to a higher proportion of initially recruited follicles that produce AMH, but fail to progress towards cyclic recruitment. The tissue
30 gene expression patterns of *MTHFR* do not show any bias towards oocyte expression. Analyzing

a sample for this mutation or other mutations (Table 1) in the *MTHFR* gene or abnormal gene expression of products of the *MTHFR* gene allows one to assess a risk of infertility.

Jeddi-Tehrani et al. (American Journal of Reproductive Immunology 66(2):149-156, 2011) investigated the effect of the *MTHFR* 677TT genotype on Recurrent Pregnancy Loss (RPL). One hundred women below 35 years of age with two successive pregnancy losses and one hundred healthy women with at least two normal pregnancies were used to assess the frequency of five candidate genetic risk factors for RPL - *MTHFR* 677C>T, *MTHFR* 1298A>C, *PAI1* -675 4G/5G (Plasminogen Activator Inhibitor-1 promoter region), *BF* -455G/A (Beta Fibrinogen promoter region), and *ITGB3* 1565T/C (Integrin Beta 3). The frequencies of the polymorphisms were calculated and compared between case and control groups. Both the *MTHFR* polymorphisms (677C>T and 1298 A>C) and the *BF* -455G/A polymorphism were found to be positively and *ITGB3* 1565T/C polymorphism was found to be negatively associated with RPL. Homozygosity but not heterozygosity for the *PAI-1* -6754G/5G polymorphism was significantly higher in patients with RPL than in the control group. The presence of both mutations of *MTHFR* genes highly increased the risk of RPL. Analyzing a sample for these mutation and other mutations (Table 1) in the *MTHFR* gene or abnormal gene expression of products of the *MTHFR* gene allows one to assess a risk of infertility.

Catechol-O-methyltransferase (COMT) In particular embodiments a mutation (472G>A) in the *COMT* gene is associated with infertility. Catechol-O-methyltransferase is known in the art to be one of several enzymes that inactivates catecholamine neurotransmitters by transferring a methyl group from SAM (S-adenosyl methionine) to the catecholamine. The AA gene variant is known to alter the enzyme's thermostability and reduces its activity 3 to 4 fold (Schmidt et al., Epidemiology 22(4): 476-485, 2011). Salih et al. (Fertility and Sterility 89(5, Supplement 1): 1414-1421, 2008) investigated the regulation of COMT expression in granulosa cells and assessed the effects of 2-ME2 (COMT product) and COMT inhibitors on DNA proliferation and steroidogenesis in JC410 porcine and HGL5 human granulosa cell lines in *in vitro* experiments. They further assessed the regulation of COMT expression by DHT (Dihydrotestosterone), insulin, and ATRA (all-trans retinoic acid). They concluded that COMT expression in granulosa cells was up-regulated by insulin, DHT, and ATRA. Further, 2-ME2 decreased, and COMT inhibition increased granulosa cell proliferation and steroidogenesis. It was hypothesized that COMT overexpression with subsequent increased level of 2-ME2 may

lead to ovulatory dysfunction. Analyzing a sample for this mutation in the *COMT* gene or abnormal gene expression of products of the *COMT* gene allows one to assess a risk of infertility.

Methionine Synthase Reductase (*MTRR*) In particular embodiments a mutation (A66G) in the Methionine Synthase Reductase (*MTRR*) gene is associated with infertility. *MTRR* is required for the proper function of the enzyme Methionine Synthase (*MTR*). *MTR* converts homocysteine to methionine, and *MTRR* activates *MTR*, thereby regulating levels of homocysteine and methionine. The maternal variant A66G has been associated with early developmental disorders such as Down's syndrome (Pozzi et al., 2009) and Spina Bifida (Doolin et al., American journal of human genetics 71(5): 1222-1226, 2002). Analyzing a sample for this mutation in the *MTRR* gene or abnormal gene expression of products of the *MTRR* gene allows one to assess the risk of infertility.

Betaine-Homocysteine S-Methyltransferase (*BHMT*) In particular embodiments a mutation (G716A) in the *BHMT* gene is associated with infertility. Betaine-Homocysteine S-Methyltransferase (*BHMT*), along with *MTRR*, assists in the Folate/B-12 dependent and choline/betaine-dependent conversions of homocysteine to methionine. High homocysteine levels have been linked to female infertility (Berker et al., Human Reproduction 24(9): 2293-2302, 2009). Benkhalifa et al. (2010) discuss that controlled ovarian hyperstimulation (COH) affects homocysteine concentration in follicular fluid. Using germinal vesicle oocytes from patients involved in IVF procedures, the study concludes that the human oocyte is able to regulate its homocysteine level via remethylation using *MTR* and *BHMT*, but not CBS (Cystathione Beta Synthase). They further emphasize that this may regulate the risk of imprinting problems during IVF procedures. Analyzing a sample for this mutation in the *BHMT* gene or abnormal gene expression of products of the *BHMT* gene allows one to assess a risk of infertility.

Ikeda et al. (Journal of Experimental Zoology Part A: Ecological Genetics and Physiology 313A(3): 129-136, 2010) examined the expression patterns of all methylation pathway enzymes in bovine oocytes and preimplantation embryos. Bovine oocytes were demonstrated to have the mRNA of *MAT1A* (Methionine adenosyltransferase), *MAT2A*, *MAT2B*, *AHCY* (S-adenosylhomocysteine hydrolase), *MTR*, *BHMT*, *SHMT1* (Serine hydroxymethyltransferase), *SHMT2*, and *MTHFR*. All these transcripts were consistently

expressed through all the developmental stages, except *MAT1A*, which was not detected from the 8-cell stage onward, and *BHMT*, which was not detected in the 8-cell stage. Furthermore, the effect of exogenous homocysteine on preimplantation development of bovine embryos was investigated *in vitro*. High concentrations of homocysteine induced hypermethylation of genomic DNA as well as developmental retardation in bovine embryos. Analyzing a sample for these irregular methylation patterns allows one to assess a risk of infertility.

Folate Receptor 2 (*FOLR2*) In particular embodiments a mutation (rs2298444) in the *FOLR2* gene is associated with infertility. Folate Receptor 2 helps transport folate (and folate derivatives) into cells. Elnakat and Ratnam (Frontiers in bioscience: a journal and virtual library 11: 506-519, 2006) implicate *FOLR2*, along with *FOLR1*, in ovarian and endometrial cancers. Analyzing sample mutations in the *FOLR2* or *FOLR1* genes or abnormal gene expression of products of the *FOLR2* or *FOLR1* genes allows one to assess a risk of infertility.

Transcobalamin 2 (*TCN2*) In particular embodiments a mutation (C776G) in the *TCN2* gene is associated with infertility. Transcobalamin 2 facilitates transport of cobalamin (Vitamin B12) into cells. Stanislawska-Sachadyn et al. (Eur J Clin Nutr 64(11): 1338-1343, 2010) assessed the relationship between *TCN2* 776C>G polymorphism and both serum B12 and total homocysteine (tHcy) levels. Genotypes from 613 men from Northern Ireland were used to show that the *TCN2* 776CC genotype was associated with lower serum B12 concentrations when compared to the 776CG and 776GG genotypes. Furthermore, vitamin B12 status was shown to influence the relationship between *TCN2* 776C>G genotype and tHcy concentrations. The *TCN2* 776C>G polymorphism may contribute to the risk of pathologies associated with low B12 and high total homocysteine phenotype. Analyzing a sample for this mutation in the *TCN2* gene or abnormal gene expression of products of the *TCN2* gene allows one to assess a risk of infertility.

Cystathionine-Beta-Synthase (*CBS*) In particular embodiments a mutation (rs234715) in the *CBS* gene is associated with infertility. With vitamin B6 as a cofactor, the Cystathionine-Beta-Synthase (*CBS*) enzyme catalyzes a reaction that permanently removes homocysteine from the methionine pathway by diverting it to the transsulfuration pathway. *CBS* gene mutations associated with decreased *CBS* activity also lead to elevated plasma homocysteine levels. Guzman et al. (2006) demonstrate that *Cbs* knockout mice are infertile. They further explain that *Cbs*-null female infertility is a consequence of uterine failure, which is a consequence of hyperhomocysteinemia or other factor(s) in the uterine environment. Analyzing a sample for this

mutation in the *CBS* gene or abnormal gene expression of products of the *CBS* gene allows one to assess a risk of infertility.

In certain embodiments, the biomarker is a genetic region that has been previously associated with female infertility. A SNP association study by targeted re-sequencing was performed to search for new genetic variants associated with female infertility. Such methods have been successful in identifying significant variants associated in a wide range of diseases (Rehman et al., 2010; Walsh et al., 2010). Briefly, a SNP association study is performed by collecting SNPs in genetic regions of interest in a number of samples and controls and then testing each of the SNPs that showed significant frequency differences between cases and controls. Significant frequency differences between cases and controls indicate that the SNP is associated with the condition of interest.

In certain embodiments, genetic loci to be investigated in a mouse model are derived from a cluster analysis, discussed below. As stated above, other methods to determine a genetic region of interest can be employed, i.e., human test results or findings published in literature.

Cluster Analysis

In addition to using infertility biomarkers identified above, methods of the invention further utilize the existing infertility knowledgebase to identify commonalities between known infertility genes and genes having no prior association with infertility. By identifying commonalities between infertility genes and genes having no prior association with infertility, one is able to expand the list of potential genes associated with infertility and guide understanding as to what gene functions and changes are causally-linked to infertility. For example, genes having commonalities with known infertility genes can be identified as potential infertility biomarkers, and used in phenotypic studies (such those performed in mice) related to infertility, thereby expanding the breadth infertility knowledgebase.

In order to determine commonalities between infertility genes and genes without prior associated with infertility, methods of the invention utilize cluster analysis techniques. Generally, a cluster analysis involves grouping a set of objects in such a way that certain objects are clustered in one group are more similar to each other than objects in another group or cluster. Methods of the invention cluster known infertility genes with genes not associated with infertility based on features such as gene expression, phenotype, and genetic pathways. From the cluster analysis, one can identify genes without prior association with infertility that exhibit

features with a high degree of similarity (relatedness) to infertility genes. Those genes exhibiting a high degree of similarity (as shown through the cluster analysis) can be identified as a potential infertility biomarker.

The following describes a clustering method used to identify a potential infertility biomarker in accordance with methods of the invention. The method is typically a computer-implemented method, e.g. utilizes a computer system that includes a processor and a computer readable storage medium. The processor of the computer system executes instructions obtained from the computer-readable storage device to perform the cluster analysis.

In accordance with to certain aspects, the method involves obtaining a gene data set that includes both known infertility genes and genes having no prior association with infertility. The genes forming the cluster data set (those associated with infertility and those not known to be associated with infertility) are typically mammalian genes. The mammalian genes may correspond to mouse genes, human, genes, or a combination thereof. A cluster analysis is then performed on the gene data set to determine a relationship between the one or more genes not associated with infertility and the known infertility genes. If a gene not associated with infertility is shown to cluster with a known infertility gene, the method provides for identifying that gene as a potential infertility biomarker. If the gene not associated with infertility does not cluster with a known infertility gene, then that gene is less likely to be causally linked to infertility in the same/similar manner as that known infertility gene.

Methods of the invention assess several features (or parameters) of genes in order to determine commonalities and thus cluster genes not associated with infertility with known infertility genes based on the commonalities. In certain embodiments, those features include gene expression, phenotypes, gene pathways, and a combination thereof. One or more of those features can contribute to a gene's position in the clustering.

Feature data (such as gene expression, phenotype, gene pathway, etc.) is obtained for both known infertility genes and genes not known to be associated with infertility. The feature and gene data is compiled to form a matrix that will be used to exhibit the cluster analysis. For example, the feature data is pre-processed to express each domain as a row and each feature as a column (or vice versa). For domains with continuous values such as gene expression, the features are the individual tissues where gene expression was measured, and each value in the matrix (X_{ij}) represents the expression of gene i in tissue j . For domains with categorical values

such as phenotypes, the features are the individual phenotypes, and each value in the matrix (X_{ij}) is a binary indicator representing whether gene i is associated with phenotype j . All of the domain specific matrices are then combined column-wise. A distance metric is then applied to each pair of rows and each pair of columns in the matrix. In certain embodiments, the distance metric is 'Distance=1-correlation'. However, it is understood that other standard distance metrics could be used (e.g. Euclidean).

Standard hierarchical clustering was then used to cluster the rows and columns of the matrix in order to determine feature commonalities between known infertility genes and other genes. Various hierarchical clustering techniques are known in the art, and can be applied to methods of the invention for clustering infertility genes with genes not associated with infertility. Hierarchical clustering techniques are described in, for example, Sturn, Alexander, John Quackenbush, and Zlatko Trajanoski. "Genesis: cluster analysis of microarray data." *Bioinformatics* 18.1 (2002): 207-208; Yeung, Ka Yee, and Walter L. Ruzzo. "Principal component analysis for clustering gene expression data." *Bioinformatics* 17.9 (2001): 763-774; Eisen, Michael B., et al. "Cluster analysis and display of genome-wide expression patterns." *Proceedings of the National Academy of Sciences* 95.25 (1998): 14863-14868. Generally, clustering involves comparing features of one or more genes not associated with features of one or more known infertility, and categorizing the genes into one or more feature groups based on the comparison. After the comparison, the cluster analysis may further involve assigning a value to the categorized genes based on a degree of relatedness. For example, genes clustered together having highly similar or the same features may be assigned a high value (e.g. positive integer). The degree of relatedness may be highlighted on the resulting cluster matrix via colors, e.g. high degree of commonality being shown in red and low degree of commonality being shown in blue.

After a hierarchical clustering technique is applied to the gene/feature data, the gene clusters are displayed against certain feature categories (e.g. phenotype/gene expression 'category'), which are then clustered to reflect commonality. For example, phenotypes of female reproduction are grouped together in one cluster, and phenotypes of embryo patterning, morphology and growth are grouped in a separate cluster, etc. The degree of relatedness or commonality between clustered genes (as determined by the cluster analysis) can then be highlighted on the resulting cluster matrix. For example, red may be used to indicate that the gene is associated with one very specific phenotype and/or is expressed at high levels in the

associated tissue/physiological system indicated on the opposite axis; whereas blue may be used to indicate that the gene is associated with a number of different and varied phenotypes and/or is expressed at low levels in the associated tissue.

By clustering genes into feature specific groups and color-coding genes with high degree of relatedness, the resulting cluster matrix of the invention advantageously allows for visualization of groups of genes that are strongly associated with phenotypes relating to particular tissues or physiological systems (i.e. clusters of interest). Thus, cluster matrices of the invention allow one to quickly identify genes without prior association with infertility as potential infertility biomarkers based on their shown association (cluster) with known infertility biomarkers. This clustering and identification of potential infertility biomarkers is done independently from and without correlating a gene's proximity with other genes within or location on the Fertilome (genomic region associated with infertility). As a result, clustering provides an additional method of identifying infertility genes of interest that can be used to complement and in addition to other techniques for identifying infertility genes of interest.

The following describes a specific example of using the above described cluster analysis to correlate genes not known to be associated with infertility and a known infertility gene.

Activin receptor 2b (ACVR2B) is a significant copy number variation identified in a cohort of patients with infertility (i.e. copy number variation in this gene was identified as being significantly associated with an infertile phenotype in humans). Activin receptor 2B is the receptor bound by Activin, a protein previously known in the art to be involved in both human and mouse reproduction and embryonic development. Activin/Nodal signaling regulates pluripotency and several aspects of patterning during early embryogenesis. Together with Inhibin and Follistatin, Activin is also involved in the complex feedback loops that selectively regulate FSH secretion.

A cluster analysis was performed that compared those features of ACVR2B and features of a plurality of genes not known to be associated with infertility. Based on the cluster analysis, several of the plurality of genes were determined to cluster with the ACVR2B gene due to a commonality between functional and phenotypic features. The genes clustered with the ACVR2B gene were thus identified as potential infertility biomarkers. FIG. 14 illustrates the results of a cluster analysis with ACVR2B.

Cluster analysis as applicable to mouse modeling is further described in more detail below. As discussed, clustering analysis provides more functional information with regards to infertility suspected genetic loci and biomarkers by putting genetic loci in clusters according to attributes including phenotype and tissue expression level/pattern. Results of the cluster analysis
5 reveal genetic loci that have a newly predicted association with the other loci in the cluster. Prior, there may have been no existing indication of a direct functional link in the literature. Thus, cluster analysis may be used to highlight new genetic loci for further phenotypic study in mouse models, and can create knowledge of how particular genetic loci cluster together to provide understanding of how mutation(s) in the gene(s) of interest might bring about the
10 molecular, cellular and physiological changes sufficient to affect particular aspects of infertility.

Attributes such as expression, phenotype, or knowledge of gene pathways or a combination of any of these can contribute to a gene's position in the clustering. Data from one, two, or any combination of these parameters are pre-processed to express each domain as a matrix with genetic loci in rows and features in columns. For domains with continuous values
15 such as gene expression, the features are the individual tissues where gene expression was measured, and each value in the matrix (X_{ij}) represents the expression of gene i in tissue j . For domains with categorical values such as phenotypes, the features are the individual phenotypes, and each value in the matrix (X_{ij}) is a binary indicator representing whether gene i is associated with phenotype j . All of the domain specific matrices are then combined column-wise. A
20 distance metric is then applied to each pair of rows and each pair of columns in the matrix (for example, 'Distance=1-correlation'), but other standard distance metrics could be used (e.g., Euclidean). Standard hierarchical clustering can then be used to cluster the rows and columns of the matrix.

The gene clusters are displayed against an attribute such as phenotype/gene expression
25 'category', which is in turn 'clustered' to reflect commonality. For example, phenotypes of female reproduction are grouped together in one cluster. Phenotypes of embryo patterning, morphology and growth are grouped in a separate cluster, etc. Measurement can be indicated by a color scale, for example, where red may indicate that the gene is associated with one very specific phenotype and/or is expressed at high levels in the associated tissue/physiological
30 system indicated on the opposite axis; whereas blue indicates the gene is associated with a number of different and varied phenotypes and/or is expressed at low levels in the associated

tissue. Therefore correlations can be visualized of groups of genetic loci that are strongly associated with phenotypes relating to particular tissues or physiological systems. The clustering is done independent of any information regarding the physical proximity of these genetic elements on the chromosome. The method of clustering allows both a narrow- and wide-scale
5 view of groups of genetic loci and their association with [a] particular phenotype(s), highlighting groups of genetic loci likely to function in a similar way and in some cases even together, to regulate particular aspects of infertility.

According to certain embodiments, a cluster analysis is created by first compiling a database is compiled that includes features attributed to each nucleotide of the human genome
10 including functional annotation such as gene boundaries, exons, splice sites, areas of putative non-coding RNAs and other elements such as promoters or CpG islands and features associated with those regions such as tissue-specific transcriptional expression from multiple mammalian systems including mouse and human, transgenic mouse strain phenotypes, mutations in genetic loci or genetic regions that have been associated with different human diseases, the relationship
15 of particular genetic loci to particular molecular or cellular pathways, gene ontology, protein-protein interactions, and mutations that have been observed. Some of the data is from public sources (e.g., mouse phenotypes) and some data is from research studies (e.g., non-public data related to mouse phenotypes and non-coding areas of interest or coding region mutations observed in patients with infertility).

20 After the database is assembled, a meta-analysis on the gene regions is performed in the following way. First, the data is pre-processing to express each domain as a matrix with genetic loci in rows and features in columns. For domains with continuous values such as gene expression, the features are the individual tissues where gene expression was measured, and each value in the matrix (X_{ij}) represents the expression of gene i in tissue j . For domains with
25 categorical values such as phenotypes, the features are the individual phenotypes, and each value in the matrix (X_{ij}) is a binary indicator representing whether gene i is associated with phenotype j . Each domain matrix has R rows and C_k columns

Each domain matrix is then scaled so that each gene has mean 0 and standard deviation 1. All of the domain specific matrices are then combined column-wise, giving a matrix with R rows
30 and $\sum C_k$ columns.

A distance metric is then applied to each pair of rows and each pair of columns in the matrix. Here, the weighted correlation value is the Pearson correlation with higher weights applied to specific features (columns). Since interest is in infertility driven clustering, infertility/reproductive associated phenotypes and tissues are given higher weights in the correlation value and hence in the distance calculation. Alternate weights could be used to emphasize other aspects of the gene information. The resulting distance value is 0 for genetic loci with identical annotation, and 1 for completely uncorrelated annotation.

Standard hierarchical clustering is then used to cluster the rows and columns of the matrix. An intensity-based coloring is used on the values in the matrix with red indicating a higher positive signal. The gene-wise distances and the associated clustering have several uses.

For example, starting from known infertility associated genetic loci in one mammalian species such as mouse, one can identify novel infertility associated genetic loci in the same species or another mammalian species that contains an orthologous gene. As an example, starting with the known human infertility gene NLRP5, Table 8 lists the most similar (smallest distance) genes to NLRP5. Most of the genes on the list have already been identified based on published studies as having an association with infertility (a validation of the approach), but several have not (e.g., ATAD2B, NR2E1). In this example, ATAD2B, NR2E1 are good candidates for studies/analysis to confirm their infertility association.

For example, starting with a partially characterized gene, impute likely phenotypes/pathways based on co-clustered genetic loci. As an example, the gene CHST8 has incomplete annotation regarding its role in human biological pathways and diseases, including infertility. Table 9 shows the genes most similar in function to CHST8 based on the clustering method. The fertility-associated genes FSHB and LHB are characterized as being similar to, or having similar function to CHST8, and are both well characterized independently. Both encode binding proteins for hormones important in female fertility. In this example, CHST8 is therefore a good candidate for studies/analysis to reveal how it is associated with infertility, for example through the disruption of the CHST8 gene in a transgenic mouse model.

For example, identify clusters of related infertility-associated genetic loci that may be used for the development of an infertility assay in humans [Pittman, Jennifer, et al. "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes." Proceedings of the National Academy of Sciences of the United States of America

101.22 (2004): 8431-8436]. Figure 14 shows a cluster of genes, each with their own particular gene annotation, curated from knowledge in the literature such as but not limited to, tissue-specific gene expression level, association of the gene or genetic region with (a) particular phenotype/s, association of the gene or genetic region with particular cellular pathway, and protein-protein interactions. Membership in a cluster is based on a genetic region demonstrating similar attributes in these domains, and on the division of the clustering tree into sections depending on the degree of functional relatedness of genetic loci within particular clusters, calculated by the attributes listed. In an alternative embodiment a method such as k-means could be used. The present methodology determines that each cluster of genetic loci may be involved with a separate aspect of fertility (e.g., oocyte development, hormone signaling, embryo implantation). These clusters could then serve as the basis of assays to assess human infertility, or as candidates for the creation of genetically altered mice to provide a model for infertility, as well as the means to test infertility treatments, such as those provided by, but not limited to, therapeutic drugs. The clusters can also be used empirically, without knowing their association with specific characteristics of infertility, by creating meta-genes. A meta-gene is a weighted combination of a set of genetic loci, and functions as a single predictor of human infertility that integrates effects from multiple similar genetic loci. The use of meta-genes can significantly increase the power of genetic/genomic studies by increasing the predictive strength and reducing the number of hypotheses tested.

Table 8

entrezGeneId	symbol	Known Infertility Association	MouseGeneId	Similarity (1-Distance)
126206	NLRP5	Y	23968	1
441161	OOEP	Y	67968	0.990508
326340	ZAR1	Y	317755	0.954272
359787	DPPA3	Y	73708	0.925278
54454	ATAD2B		320817	0.768295
8115	TCL1A	Y	21432	0.729399
4361	MRE11A		17535	0.728909
4360	MRC1	Y	17533	0.727167
7101	NR2E1		21907	0.719154

23633	KPNA6		16650	0.712841
2827	GPR3	Y	14748	0.709265
7783	ZP2	Y	22787	0.709177
200424	TET3		194388	0.707759
127343	DMBX1	Y	140477	0.704141
10361	NPM2	Y	328440	0.700169
7784	ZP3	Y	22788	0.696949
9210	BMP15	Y	12155	0.688272
22917	ZP1	Y	22786	0.688209
54014	BRWD1	Y	93871	0.681323
344018	FIGLA	Y	26910	0.674247
6533	SLC6A6	Y	21366	0.673478
2661	GDF9	Y	14566	0.664854
27252	KLHL20		226541	0.662994
204801	NLRP11	Y		0.655971
654790	PCP4L1	Y	66425	0.655923

Table 9

entrezGeneId	symbol	Known Infertility Association	MouseGeneId	Similarity (1-Distance)
64377	CHST8		68947	1
2488	FSHB	Y	14308	0.807603
3972	LHB	Y	16866	0.799529
8022	LHX3	Y	16871	0.720396
23373	CRTC1	Y	382056	0.68314
2798	GNRHR	Y	14715	0.680513
7425	VGf	Y	381677	0.673726
54551	MAGEL2		27385	0.656467
1813	DRD2	Y	13489	0.650742

5617	PRL	Y	19109	0.643812
1081	CGA	Y	12640	0.62561
5122	PCSK1		18548	0.624284
3763	KCNJ6		16522	0.624099
6447	SCG5	Y	20394	0.611227
6833	ABCC8	Y	20927	0.602154
9985	REC8	Y	56739	0.592734
273	AMPH		218038	0.592075
2688	GH1		14599	0.587602
4438	MSH4	Y	55993	0.571955
113091	PTH2		114640	0.559548
11144	DMC1	Y	13404	0.55841
25970	SH2B1	Y	20399	0.55654
6658	SOX3	Y	20675	0.553021
135935	NOBOX	Y	18291	0.551976
3990	LIPC		15450	0.550449

In an aspect of the invention, genetic loci are ranked according to their expression levels in humans and mice. For example, it is determined whether a biomarker is expressed in mice. If the biomarker is expressed in mice, the biomarker receives a higher ranking. If the biomarker is also expressed in humans, the biomarker is ranked even higher by the ranking system. If a biomarker is not expressed in mice, or in humans, it would receive a low ranking. A biomarker would receive the lowest ranking if it was expressed neither in mouse nor in human. Known methods in the art can be employed to rank genetic regions. It should be appreciated that any known ranking methodology can be utilized in the present invention, as discussed above. For example, the Friedman test, Kruskal-Wallis test, Spearman's rank correlation coefficient, Wilcoxon rank-sum test, and/or Wilcoxon signed-rank test are known statistical methods. The Friedman test is similar to the parametric repeated measures ANOVA; it is used to detect differences in treatments across multiple test attempts. The procedure involves ranking each row (or block) together, then considering the values of ranks by columns. See Friedman, Milton

(December 1937). "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". *Journal of the American Statistical Association* (American Statistical Association) 32 (200): 675–701. Also, the Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient measures the strength of association between two ranked variables. See Lehman, Ann (2005). *Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide*. Cary, NC: SAS Press. p. 123. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e., it is a paired difference test). See Wilcoxon, Frank (Dec 1945). "Individual comparisons by ranking methods". *Biometrics Bulletin* 1 (6): 80–83.

In an aspect of the invention, another possible ranking scheme employs listing genes in order from most to least statistically significant, when the correlation with phenotype in mice is determined. In this method, confidence intervals and p values are employed, where P-values < .025 are considered statistically significant. A series of linear regression models are fit, where the outcome variable is the phenotype expression score for a given gene, and the independent variables are group (expressed phenotype v. control) and principal component derived ethnicity (for humans) or strain (for mice) (continuous). The p-value for group is used for statistical inference. The model is fit once for each gene.

In an aspect of the invention, another possible gene ranking scheme, genetic loci are ranked according to a Celmatix FertiloScore™, G1Version2, that reflects the likelihood that a gene is involved in fertility or reproduction. This score is computed using a database of mined and curated data, containing attributes for each gene in the genome. These attributes include: diseases and disorders related to infertility, molecular pathways, molecular interactions, gene clusters, mouse phenotypes associated with each gene, gene expression data in reproductive tissues, proteomics data in oocytes, and accrued information from scientific publications through text-mining.

The process for ranking fertility-related attributes of a gene or genetic region (locus) to obtain a score is carried out by the SESMe algorithm. The SESMe algorithm is applied to a database of features and attributes that might make a particular gene important for fertility. The algorithm assigns a score and a relative weight to each feature to then rank genetic regions from

most to least important (or vice versa) by weighting features and attributes associated with that genetic region. For example, a score is assigned to a gene by compiling the combined weighted values of attributes associated with that gene. After each gene is scored based on its weighted attributes, the genetic loci can be ranked in order of importance in accordance with their score.

- 5 The weighted value for each infertility attribute may be scaled in any manner including and not limited to assigning a positive or negative integer to reflect the significance or severity of the attribute to infertility.

In certain embodiments, the weighted value for gene infertility attributes may be on a scale from -10 to +10. A +10 may indicate that an attribute of a gene being scored is highly associated with infertility because that attribute is prevalently found in infertile patient populations. A +4 may represent an attribute that is a latent infertility marker, meaning it will not cause infertility on its own, but may lead to infertility upon influence of external factors such as aging and smoking. Whereas +2 may represent an attribute found in some infertile patients but nothing directly relates the attribute to infertility. A zero on the scale may include an attribute not yet known to have any effect or any negative effect towards infertility. A -10 may include an attribute shown not to affect infertility whatsoever. Further, embodiments provide for the weighted scale to include a +1 for attributes that are commonly found in infertile patient populations, 0.5 for attributes similar to those found in infertile patient populations, and 0 for attributes without a causal link to infertility.

20 In addition, weighted values for attributes may be normalized based on the known significance of that attribute towards infertility. For example and in certain embodiments, when scoring attributes of a particular gene, each attribute may be assigned a 0 if the attribute is absent and a 1 if the attribute is present. The attributes may then be normalized based on the infertility significance of that attribute. For example, if the attribute is a genetic mutation known to be associated with infertility, then that attribute may be normalized by a factor of 5. In another example, if the attribute is a signaling pathway defect sometimes associated with infertility, then that attribute may be normalized by a factor of 2.

In an aspect of the invention, another possible gene ranking scheme involves the relative degree of infertility, subfertility, or premature decline in fertility risk associated with novel or common mutations or variants in a fertility gene. Genetic loci are ranked according to a Celmatix Fertilome™Score, G1Version3, that reflects the likelihood a gene is involved in

fertility or reproduction. This score is computed using a database of mined and curated data, containing attributes for each gene in the genome. These attributes include: diseases and disorders related to infertility, molecular pathways, molecular interactions, gene clusters, mouse phenotypes associated with each gene, gene expression data in reproductive tissues, proteomics data in oocytes, and accrued information from scientific publications through text-mining. The Celmatix Fertilome™Score G1Version3 differs from G1Version2 because it contains more fertility genetic loci as an input for the score calculation.

Mouse Model

The ability to engineer the mouse genome has proven useful for a variety of applications in research, medicine and biotechnology. Transgenic mice have become powerful reagents for modeling genetic disorders, understanding embryonic development and evaluating therapeutics. These mice and the cell lines derived from them have also accelerated basic research by allowing scientists to assign functions to genetic loci, dissect genetic pathways, and manipulate the cellular or biochemical properties of proteins.

Generation of a mouse model may be accomplished by any known method in the art. This can involve, but is not limited to, the addition of exogenous sequences of DNA to the genome of an animal during its earliest stage of development (the zygote) to permanently and heritably alter the expression of a particular gene or group of loci's expression. Methodologically, this can involve, but is not limited to, the pronuclear injection of short sequences of oligonucleotides derived *in vitro*, which replace endogeneous DNA sequences through homologous recombination and can therefore be designed to encode for mutated versions of genes or genetic regions. The generation of mouse models can also include, but is not limited to, the insertion of DNA sequences (designed to be expressed at an enhanced or attenuated level when compared to that of their endogenous copy) into retroviral vectors that allow the DNA sequences to replace their endogenous (normal) copy in the genome. See for example, Bedell, M. A., et al. Mouse models of human disease. Part I: Techniques and resources for genetic analysis in mice. *Genes and Development* 11, 1-10 (1997a); Rosenthal, N., & Brown, S. The mouse ascending: Perspectives for human-disease models. *Nature Cell Biology* 9, 993-999 (2007) doi:10.1038/ncb437; Yang, S. H. et al. Towards a transgenic model of Huntington's disease in a non-human primate. *Nature* 453, 921-924 (2008) doi:10.1038/nature06975; Yu, Y., & Bradley,

A. Mouse genomic technologies: Engineering chromosomal rearrangements in mice. *Nature Reviews Genetics* 2, 780-790 (2001).

Using any or all of these methods, many different types of mutations can be introduced into any particular genetic region, including null or point mutations and complex chromosomal rearrangements such as large deletions, translocations, or inversions (Bedell et al., 1997a). Depending on the mutation introduced into the animal and as understood in the art, the genetically modified animal may be referred to as a “knockin” or “knockout” animal, or the mutation itself may be referred to as a “knockin” mutation or “knockout” mutation.

Methods that target a particular genetic region for alteration in expression are particularly useful if a single gene is shown to be the primary cause of a disease., and indeed more than 3,000 genes have been targeted and altered in mice. Most of the targeted and altered genes have been related to disease (Hardouin & Nagy, 2000). Many genetically altered mice have similar, if not identical, phenotypes to human patients with lesions in the same/related genetic regions. Many mouse models therefore represent useful tools with which to model human disease.

In an aspect of the invention therefore, genetic loci that are identified as being highly ranked in association with particular aspects of infertility or reproductive biology and have previously never been directly associated with those characteristics in humans or in mice, would serve as good candidates for the generation of mouse models for infertility. These mouse models would in turn provide tools for testing therapeutic agents designed to overcome certain aspects of infertility related to particular molecular aetiologies.

Testing of Therapeutic Agents

The genetically altered mouse is then assessed to determine whether the gene or biomarker expresses a phenotype. Genetically-altered test animals that show presence of an infertility phenotype are useful for therapeutic testing. For example, a genetically altered mouse expressing a phenotype can be dosed or exposed to a therapeutic agent such as, Human Chorionic Gonadotropin (hCG), (such as Pregnyl, Novarel, Ovidrel, and Profasi); Follicle Stimulating Hormone (FSH), (such as Follistim, Fertinex, Bravelle, and Gonal-F); Human Menopausal Gonadotropin (hMG), (such as Pergonal, Repronex, and Metrodin) or Gonadotropin Releasing Hormone (GnRH), (such as Factrel and Lutrepulse); Gonadotropin Releasing Hormone Agonist (GnRH agonist), (such as Lupron, Zoladex, and Synarel); or Gonadotropin

Releasing Hormone Antagonist (GnRH antagonist), (such as Antagon and Cetrotide) to determine if the therapeutic agent is effective at overcoming infertility. A therapeutic agent that rescues the phenotype, i.e., returns or partially re-establishes the wild type fertility phenotype, is a good drug candidate.

5

Predictive Value

Infertility may not be the result of a single genomic alteration, but rather may be the result of a combination of multiple factors or multiple alterations. Methods of the invention provide a better understanding of the molecular pathways underlying human fertility. For example, presence of an infertility-associated phenotype is used as a factor in ranking the importance of a gene in a database of genes associated with infertility in humans by associated the gene (or more often a mutation) with the phenotype. A correlation between the presence of an allele or a mutation in a gene with phenotype increases or decreases the predictive value of the contribution of the genomic region to phenotype.

15

Computer Systems

FIG. 15 illustrates a computer system 401 useful for implementing methodologies described herein. A system of the invention may include any one or any number of the components shown in FIG. 15. Generally, a system 401 may include a computer 433 and a server computer 409 capable of communication with one another over network 415. Additionally, data may optionally be obtained from a database 405 (e.g., local or remote). In some embodiments, systems include an instrument 455 for obtaining sequencing data, which may be coupled to a sequencer computer 451 for initial processing of sequence reads.

20

In some embodiments, methods are performed by parallel processing and server 409 includes a plurality of processors with a parallel architecture, i.e., a distributed network of processors and storage capable of collecting, filtering, processing, analyzing, ranking genetic data obtained through methods of the invention. The system may include a plurality of processors configured to, for example, 1) collect genetic data from different modalities: a) one or more infertility databases 405 (e.g. infertility databases, including private and public fertility-related data), b) from one or more sequencers 455 or sequencing computers 451, c) from mouse modeling, etc; 2) filter the genetic data to identify genetic variations; 3) associate genetic

30

variations with infertility using methods described throughout the application (e.g., filtering, clustering, etc.); 4) determine statistical significance of genetic variations based on fertility criteria defined herein (e.g., Example 18); and 5) characterize/identify the genetic variations as infertility biomarkers.

5 By leveraging genetic data sets obtained across different sources, applying layers of analyses (i.e., filtering, clustering, etc.) to genetic data, and quantifying/qualifying statistical significance of that genetic data, systems of the invention are able yield and identify new infertility biomarkers that previously could not be determined to have any association with infertility. For example, methods of the invention utilize data sets of different modalities. The data sets range include data obtained from infertility databases (e.g., public and private), sequencing data (e.g., whole genome sequencing from one or more biological samples), and genetic data obtained from mouse modeling, etc. Several layers of analysis are then applied to the genetic data to identify whether variations are potentially associated with infertility. Particularly, the genetic data sets are subject to evolutionary conservation analysis, filtering analysis (see FIG. 5) and/or subject to clustering analysis. After those analyses are applied, the variants potentially associated with infertility are then assessed for biological and statistical significance. The variants that are determined to be statistically significant are then classified as infertility biomarkers, even if those variant had no prior association with infertility. Accordingly, using the invention's multi-modal and layered analysis, one is able to identify infertility biomarkers that would not have been identified or associated with infertility using standard techniques (i.e. comparing genetic sequences of an abnormal, infertile population to genetic sequences of a normal, fertile population).

While other hybrid configurations are possible, the main memory in a parallel computer is typically either shared between all processing elements in a single address space, or distributed, i.e., each processing element has its own local address space. (Distributed memory refers to the fact that the memory is logically distributed, but often implies that it is physically distributed as well.) Distributed shared memory and memory virtualization combine the two approaches, where the processing element has its own local memory and access to the memory on non-local processors. Accesses to local memory are typically faster than accesses to non-local memory.

Computer architectures in which each element of main memory can be accessed with equal latency and bandwidth are known as Uniform Memory Access (UMA) systems. Typically, that can be achieved only by a shared memory system, in which the memory is not physically distributed. A system that does not have this property is known as a Non-Uniform Memory
5 Access (NUMA) architecture. Distributed memory systems have non-uniform memory access.

Processor–processor and processor–memory communication can be implemented in hardware in several ways, including via shared (either multiported or multiplexed) memory, a crossbar switch, a shared bus or an interconnect network of a myriad of topologies including star, ring, tree, hypercube, fat hypercube (a hypercube with more than one processor at a node), or n-
10 dimensional mesh.

Parallel computers based on interconnected networks must incorporate routing to enable the passing of messages between nodes that are not directly connected. The medium used for communication between the processors is likely to be hierarchical in large multiprocessor machines. Such resources are commercially available for purchase for dedicated use, or these
15 resources can be accessed via “the cloud,” e.g., Amazon Cloud Computing.

A computer generally includes a processor coupled to a memory and an input-output (I/O) mechanism via a bus. Memory can include RAM or ROM and preferably includes at least one tangible, non-transitory medium storing instructions executable to cause the system to perform functions described herein. As one skilled in the art would recognize as necessary or
20 best-suited for performance of the methods of the invention, systems of the invention include one or more processors (e.g., a central processing unit (CPU), a graphics processing unit (GPU), etc.), computer-readable storage devices (e.g., main memory, static memory, etc.), or combinations thereof which communicate with each other via a bus.

A processor may be any suitable processor known in the art, such as the processor sold
25 under the trademark XEON E7 by Intel (Santa Clara, CA) or the processor sold under the trademark OPTERON 6200 by AMD (Sunnyvale, CA).

Input/output devices according to the invention may include a video display unit (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT) monitor), an alphanumeric input device (e.g., a keyboard), a cursor control device (e.g., a mouse or trackpad), a disk drive unit, a signal
30 generation device (e.g., a speaker), a touchscreen, an accelerometer, a microphone, a cellular

radio frequency antenna, and a network interface device, which can be, for example, a network interface card (NIC), Wi-Fi card, or cellular modem.

EXAMPLES

5

Example 1 - Identification of oocyte proteins

Oocytes are collected from females, for example mice, by superovulation, and zona pellucidae are removed by treatment with acid Tyrode solution. Oocyte plasma membrane (oolemma) proteins exposed on the surface can be distinguished at this point by biotin labeling.

10 The treated oocytes are washed in 0.01 M PBS and treated with lysis buffer (7 M urea, 2 M thiourea, 4% (w/v) 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), 65 mM dithiothreitol (DTT), and 1% (v/v) protease inhibitor at -80°C). Oocyte proteins are resolved by one-dimensional or two-dimensional SDS-PAGE. The gels are stained, visualized, and sliced. Proteins in the gel pieces are digested (12.5 ng/μl trypsin in 50 mM ammonium bicarbonate
15 overnight at 37°C), and the peptides are extracted and microsequenced.

Example 2 - Sample Population for Identification of Infertility-Related Polymorphisms

Genomic DNA is collected from 30 female subjects (15 who have failed multiple rounds of IVF versus 15 who were successful). In particular, all of the subjects are under age 38.

20 Members of the control group succeeded in conceiving through IVF. Members of the test group have a clinical diagnosis of idiopathic infertility, and have failed three or more rounds of IVF with no prior pregnancy. The women are able to produce eggs for IVF and have a reproductively normal male partner. To focus on infertility resulting from oocyte defects (and eliminate factors such as implantation defects) women who have subsequently conceived by egg donation are
25 favored.

Example 3 - Sample Population for Identification of Infertility-Related Polymorphisms

In a follow-up study of a larger cohort, genomic DNA is collected from 300 female subjects (divided into groups having profiles similar to the groups described above).

30 The DNA sequence polymorphisms to be investigated are selected based on the results of small initial studies.

Example 4 - Sample Population for Identification of Premature Ovarian Failure (POF) and Premature Maternal Aging Polymorphisms

Genomic DNA is collected from 30 female subjects who are experiencing symptoms of
5 premature decline in egg quality and reserve including abnormal menstrual cycles or
amenorrhea. In particular, all of the subjects are between the ages of 15-40 and have follicle
stimulating hormone (FSH) levels of over 20 international units (IU) and a basal antral follicle
count of under 5. Members of the control group succeeded in conceiving through IVF. Members
of the test group have no previous history of toxic exposure to known fertility damaging
10 treatments such as chemotherapy. Members of this group may also have one or more female
family member who experienced menopause before the age of 40.

Example 5 - Sample Procurement and Preparation

Blood is drawn from patients at fertility clinics for standard procedures such as
15 gauging hormone levels and many clinics bank this material after consent for future research
projects. Although DNA is easily obtained from blood, wider population sampling is
accomplished using home-based, noninvasive methods of DNA collection such as saliva
using an Oragene DNA self collection kit (DNA Genotek).

Blood samples - Three-milliliter whole blood samples are venously collected
20 and treated with sodium citrate anticoagulant and stored at 4 °C until DNA extraction.

Whole Saliva - Whole saliva is collected using the Oragene DNA selfcollection
kit following the manufacturer's instructions. Participants are asked to rub their
tongues around the inside of their mouths for about 15 sec and then deposit approximately 2
ml saliva into the collection cup. The collection cup is designed so that the solution from the
25 vial's lower compartment is released and mixes with the saliva when the cap is securely
fastened. This starts the initial phase of DNA isolation, and stabilizes the saliva sample for
long-term storage at room temperature or in low temperature freezers. Whole saliva samples
are stored and shipped, if necessary, at room temperature. Whole saliva has the potential
advantage over other non-invasive DNA sampling methods, such as buccal and oral rinse, of
30 providing large numbers of nucleated cells (eg., epithelial cells, leukocytes) per sample.

Blood clots – Clotted blood that is usually discarded after extraction through serum

separation, for other laboratory tests such as for monitoring reproductive hormone levels is collected and stored at -80 °C until extraction.

Sample Preparation - Genomic DNA is prepared from patient blood or saliva for downstream sequencing applications with commercially available kits (e.g., Invitrogen's ChargeSwitch® gDNA Blood Kit or DNA Genotek kits, respectively). Genomic DNA from clotted is prepared by standard methods involving proteinase K digestion, salt/chloroform extraction and 90% ethanol precipitation of DNA. (see N Kanai et al., 1994, "Rapid and simple method for preparation of genomic DNA from easily obtainable clotted blood," J Clin Pathol 47:1043-1044, which is incorporated by reference in its entirety for all purposes).

Example 6 - Manufacturing of a Customized Oligonucleotide Library

A customized oligonucleotide library can be used to enrich samples for DNAs of interest. Several methods for manufacturing customized oligonucleotide libraries are known in the art. In one example, Nimblegen sequence capture custom array design is used to create a customized target enrichment system tailored to infertility related genetic loci. A customized library of oligonucleotides is designed to target genetic regions of Tables 1-7. The custom DNA oligonucleotides are synthesized on a high density DNA Nimblegen Sequence Capture Array with Maskless Array Synthesizer (MAS) technology. The Nimblegen Sequence Capture Array system workflow is array based and is performed on glass slides with an X1 mixer (Roche NimbleGen) and the NimbleGen Hybridization System.

In a similar example, Agilent's eArray (a web-based design tool) is used to create a customized target enrichment system tailored to infertility related genetic loci. The SureSelect Target Enrichment System workflow is solution-based and is performed in microcentrifuge tubes or microtiter plates. A customized oligonucleotide library is used to enrich samples for DNA of interest. Agilent's eArray (a web-based design tool) is used to create a customized target enrichment system tailored to infertility related genetic loci. A customized library is designed to target genetic regions of Tables 1-7. The custom RNA oligonucleotides, or baits, are biotinylated for easy capture onto streptavidin-labeled magnetic beads and used in Agilent's SureSelect Target Enrichment System. The SureSelect Target Enrichment System workflow is solution-based and is performed in microcentrifuge tubes or microtiter plates.

Example 7 - Capture of Genomic DNA

Genomic DNA is sheared and assembled into a library format specific to the sequencing instrument utilized downstream. Size selection is performed on the sheared DNA and confirmed by electrophoresis or other size detection method.

Several methods to capture genomic DNA are known in the art. In one example, the size-selected DNA is purified and the ends are ligated to annealed oligonucleotide linkers from Illumina to prepare a DNA library. DNA-adaptor ligated fragments are hybridized to a Nimblegen Sequence Capture array using an X1 mixer (Roche NimbleGen) and the Roche NimbleGen Hybridization System. After hybridization, are washed and DNA fragments bound to the array are eluted with elution buffer. The captured DNA is then dried by centrifugation, rehydrated and PCR amplified with polymerase. Enrichment of DNA can be assessed by quantitative PCR comparison to the same sample prior to hybridization.

In a similar example, the size-selected DNA is incubated with biotinylated RNA oligonucleotides “baits” for 24 hours. The RNA/DNA hybrids are immobilized to streptavidin-labeled magnetic beads, which are captured magnetically. The RNA baits are then digested, leaving only the target selected DNA of interest, which is then amplified and sequenced.

Example 8 - Sequencing of Target Selected DNA

Target-selected DNA is sequenced by a paired end (50bp) re-sequencing procedure using Illumina.’s Genome Analyzer. The combined DNS targeting and resequencing provides 45 fold redundancy which is greater than the accepted industry standard for SNP discovery.

Example 9 – Correlation of Polymorphisms with Fertility

Polymorphisms among the sequences of target selected DNA from the pool of test subjects are identified, and may be classified according to where they occur in promoters, splice sites, or coding regions of a gene. Polymorphisms can also occur in regions that have no apparent function, such as introns and upstream or downstream non-coding regions. Although such polymorphisms may not be informative as to the functional defect of an allele, nevertheless, they are linked to the defect and useful for predicting infertility. The polymorphisms are analyzed statistically to determine their correlation with the fertility status

of the test subjects. The statistical analysis indicates that certain polymorphisms identify gene defects that by themselves (homozygous or heterozygous) are sufficient to cause infertility. Other polymorphisms identify genetic variants that reduce, but do not eliminate fertility. Other polymorphisms identify genetic variants that have an apparent effect on fertility only in the presence of particular variants of other genetic loci. Other polymorphisms identify genetic variants that have an apparent effect on fertility only in the presence of particular phenotypes. Other polymorphisms identify genetic variants that have an apparent effect on fertility only in the presence of particular environmental exposures. Still other polymorphisms identify genetic variants that have an apparent effect on fertility only in the presence of any combination of particular variants of other genetic loci, presence of particular phenotypes, and particular environmental exposures.

Example 10 – Correlation of Polymorphisms with Premature Ovarian Failure (POF)

Polymorphisms among the sequences of target selected DNA from the pool of test subjects are identified, and may be classified according to where they occur in promoters, splice sites, or coding regions of a gene. Polymorphisms can also occur in regions that have no apparent function, such as introns and upstream or downstream non-coding regions. Although such polymorphisms may not be informative as to the functional defect of an allele, nevertheless, they are linked to the defect and useful for predicting likelihood of premature ovarian failure (POF). The polymorphisms are analyzed statistically to determine their correlation with the POF status of the test subjects. The statistical analysis indicates that certain polymorphisms identify gene defects that by themselves (homozygous or heterozygous) are sufficient to cause POF. Other polymorphisms identify genetic variants that increase the likelihood, but do not cause POF. Other polymorphisms identify genetic variants that have an apparent effect on POF only in the presence of particular variants of other genetic loci. Other polymorphisms identify genetic variants that have an apparent effect on POF only in the presence of particular phenotypes. Other polymorphisms identify genetic variants that have an apparent effect on POF only in the presence of particular environmental exposures. Still other polymorphisms identify genetic variants that have an apparent effect on POF only in the presence of any combination of particular variants of other genetic loci, presence of particular phenotypes, and particular environmental exposures.

Example 11 – Correlation of Polymorphisms with Premature Maternal Aging

Polymorphisms among the sequences of target selected DNA from the pool of test subjects are identified, and may be classified according to where they occur in promoters, splice sites, or coding regions of a gene. Polymorphisms can also occur in regions that have no apparent function, such as introns and upstream or downstream non-coding regions. Although such polymorphisms may not be informative as to the functional defect of an allele, nevertheless, they are linked to the defect and useful for predicting likelihood of premature decline in ovarian reserve and egg quality (i.e., maternal aging). The polymorphisms are analyzed statistically to determine their correlation with the maternal aging status of the test subjects. The statistical analysis indicates that certain polymorphisms identify gene defects that by themselves (homozygous or heterozygous) are sufficient to cause premature maternal aging. Other polymorphisms identify genetic variants that increase the likelihood, but do not cause premature maternal aging. Other polymorphisms identify genetic variants that have an apparent effect on premature maternal aging only in the presence of particular variants of other genetic loci. Other polymorphisms identify genetic variants that have an apparent effect on premature maternal aging only in the presence of particular phenotypes. Other polymorphisms identify genetic variants that have an apparent effect on premature maternal aging only in the presence of particular environmental exposures. Still other polymorphisms identify genetic variants that have an apparent effect on premature maternal aging only in the presence of any combination of particular variants of other genetic loci, presence of particular phenotypes, and particular environmental exposures.

Example 12 – Diagnostics and Counseling

A library of nucleic acids in an array format is provided for infertility diagnosis. The library consists of selected nucleic acids for enrichment of genetic targets wherein polymorphisms in the targets are correlated with variations in fertility. A patient nucleic acid sample (appropriately cleaved and size selected) is applied to the array, and patient nucleic acids that are not immobilized are washed away. The immobilized nucleic acids of interest are then eluted and sequenced to detect polymorphisms. According to the polymorphisms detected, the fertility status of the patient is evaluated and/or quantified. The patient is accordingly advised as

to the suitability and likelihood of success of a fertility treatment or suitability or necessity of a particular *in vitro* fertilization procedure.

Example 13 – Diagnostics and Counseling

5 A complete DNA sequence of any number of or all of the genes in Tables 1-7 is determined using a targeted resequencing protocol. According to the polymorphisms detected and the phenotypic traits and environmental exposures reported, the fertility status of the patient is evaluated and/or quantified. The patient is accordingly advised as to the suitability and likelihood of success of a fertility treatment or suitability or necessity of a particular *in vitro*
10 fertilization procedure.

Example 14 – Diagnostics and Counseling

 A library of nucleic acids in an array format is provided for infertility diagnosis. The library consists of selected nucleic acids for enrichment of genetic targets wherein
15 polymorphisms in the targets are correlated with variations in fertility. A patient nucleic acid sample (appropriately cleaved and size selected) is applied to the array, and patient nucleic acids that are not immobilized are washed away. The immobilized nucleic acids of interest are then eluted and sequenced to detect polymorphisms. According to the polymorphisms detected and the phenotypic traits and environmental exposures reported, the POF status of the patient or
20 likelihood of future POF occurrence is evaluated and/or quantified. The patient is accordingly advised as to whether preventative egg or ovary preservation is indicated.

Example 15 – Diagnostics and Counseling

 A complete DNA sequence of any number of or all of the genes in Tables 1-7 is
25 determined using a targeted resequencing protocol. According to the polymorphisms detected and the phenotype and environmental exposures reported, the fertility status of the patient is evaluated and/or quantified. According to the polymorphisms detected and the phenotypic traits and environmental exposures reported, the POF status of the patient or likelihood of future POF occurrence is evaluated and/or quantified. The patient is accordingly advised as to whether
30 preventative egg or ovary preservation is indicated.

Example 16 – Diagnostics and Counseling

A library of nucleic acids in an array format is provided for infertility diagnosis. The library consists of selected nucleic acids for enrichment of genetic targets wherein polymorphisms in the targets are correlated with variations in fertility. A patient nucleic acid sample (appropriately cleaved and size selected) is applied to the array, and patient nucleic acids that are not immobilized are washed away. The immobilized nucleic acids of interest are then eluted and sequenced to detect polymorphisms. According to the polymorphisms detected and the phenotypic traits and environmental exposures reported, the maternal aging status of the patient or likelihood of future premature maternal aging occurrence is evaluated and/or quantified. The patient is accordingly advised as to whether preventative egg or ovary preservation, minimization of certain environmental exposures such as alcohol intake or smoking, or mitigation of certain phenotypes such as having children at a younger age is indicated.

Example 17 – Diagnostics and Counseling

A complete DNA sequence of any number of or all of the genes in Tables 1-7 is determined using a targeted resequencing protocol. According to the polymorphisms detected and the phenotypic traits and environmental exposures reported, the fertility status of the patient is evaluated and/or quantified. According to the polymorphisms detected and the phenotype and environmental exposures reported, the maternal aging status of the patient or likelihood of future premature maternal aging occurrence is evaluated and/or quantified. The patient is accordingly advised as to whether preventative egg or ovary preservation, minimization of certain environmental exposures such as alcohol intake or smoking, or mitigation of certain phenotypes such as having children at a younger age is indicated.

Example 18 – Whole Genome Sequencing for Female Infertility Biomarker Discovery

Whole genome sequencing (WGS) allows one to characterize the complete nucleic acid sequence of an individual's genome. With the amount of data obtained from WGS, a comprehensive collection of an individual's genetic variation is obtainable, which provides great potential for genetic biomarker discovery. The data obtained from WGS can be advantageously used to expand the ability to identify and characterize female infertility biomarkers. However,

the ability to identify unknown variations of fertility significance within the vast WGS datasets is a challenging task that is analogous to finding a needle in a haystack.

Methods of the invention, according to certain embodiments, rely on bioinformatics to filter through WGS data in order to identify and prioritize variations of infertility significance.

5 Specifically, the invention relies on a combination of clinical phenotypic data and an infertility knowledgebase to rank and/or score genomic regions of interest and their likely impact on different fertility disorders. In certain aspects, the filtering approach involves assessing sequencing data to identify genomic variations, identifying at least one of the variations as being in a genomic region associated with infertility, determining whether the at least one variation is a
10 biologically-significant variation and/or a statistically-significant variation, and characterizing at least one identified variation as an infertility biomarker based on the determining step. A genomic region associated with infertility is any DNA sequence in which variation is associated with a change in fertility. Such regions may include genes (e.g., any region of DNA encoding a functional product), genetic regions (e.g., regions including genes and intergenic regions with a particular focus on regions conserved throughout evolution in placental mammals), and gene
15 products (e.g., RNA and protein). In particular embodiments, the infertility-associated genetic region is a maternal effect gene, as described above. In particular embodiments, the infertility-associated genetic region is a gene (including exons, introns, and evolutionarily conserved regions of DNA flanking either side of said gene) that impacts fertility.

20 This filtering approach facilitates rapid identification of functionally relevant variants within genomic regions of significance for fertility. The identified variations with infertility significance obtained from WGS data may be used in diagnostic testing, and ultimately assist physicians in data interpretation, guide fertility therapeutics, and clarify why some patients are not responding to treatment. The following illustrates use of WGS data to identify variants of
25 interest in accordance with methods of the invention.

FIG. 5 generally illustrates filtering through variations obtained from WGS sequencing data in order to identify variations of infertility significance. As shown in FIG. 5, the first step is to identify sequence variants in whole genome sequence. A typical whole genome can include up to four million variants. The next filtering step involves eliminating variants outside of
30 regions of interest for female fertility (which amounts to about one million variants). Next, the filtering method isolates variants within regions of interest for female fertility, which is described

herein as Fertilome nucleic acid (i.e., regions of the human genome that control egg quality and fertility). Variations located within the Fertilome nucleic acid may be in the 100,000s. The variations within the Fertilome nucleic acid are further filtered to identify and score variations of infertility significance (such variations are typically present in double digits). Particularly,

5 variations of infertility significance include those within regions predicted to effect biological function or that show a statistical correlation to infertility or treatment failure.

Biologically-significant variations within the Fertilome nucleic acid include mutations that result in a change: 1) to a different amino acid predicted to alter the folding and/or structure of the encoded protein, 2) to a different amino acid occurring at a site with high evolutionarily

10 conservation in mammals, 3) that introduces a premature stop termination signal, 4) that causes a stop termination signal to be lost, 5) that introduces a new start codon, 6) that causes a start codon to be lost or 7) that disrupts a splicing signal. Statistically-significant variations within the Fertilome nucleic acid are described in relation to and listed in Tables 2 and 3. Other methods for classifying variations as statistically- or biologically- significant includes scoring variations

15 using an infertility knowledgebase (which is described in relation to Tables 5-7 above and FIG. 6 below). The infertility knowledgebase ranks genetic loci based on attributes associated with infertility. The attributes include: diseases and disorders related to infertility, molecular pathways, molecular interactions, gene clusters, mouse phenotypes associated with each gene, gene expression data in reproductive tissues, proteomics data in oocytes, and accrued

20 information from scientific publications through text-mining. List of ranked genes of interest are provided in Tables 5-7.

FIG. 6 illustrates various data sources integrated into the infertility knowledgebase for analyzing whole-genome sequencing data according to certain embodiments. As shown in FIG. 6, information is obtained from private and public fertility-related data. Private and/or public

25 fertility-related data may include genetic loci that regulate processes of implantation, idiopathic infertility genetic loci, polycystic ovary syndrome (PCOS) genetic loci, egg quality genetic loci, endometriosis genetic loci, and premature ovarian failure genetic loci. The private and/or public fertility-related data is then subjected to the ABCoRE Algorithm to provide genomic regions and variations of interest that can be introduced into a fertility database evidence matrix along with

30 other fertility-related information. As described in the detailed description, the ABCoRE algorithm identifies fertility regions of interest by performing evolutionary conservation analysis

of one or more genetic loci obtained from the private and/or public fertility-related data. The other fertility-related information includes, for example, protein-protein interactions, pathway interactions, gene orthologs and paralogs, genomic “hotspots”, gene protein expression and meta-analysis, and data from genomic studies. In operation, whole genomic sequencing data is compared to the compiled data in the fertility database evidence matrix to facilitate identification of potential genetic regions important for fertility. The fertility database evidence matrix filters through WGS variants to identify variants of fertility significance. In certain embodiments, the whole genomic sequencing data is also subjected to the SESMe algorithm that ranks each genetic region from most to least important for different aspects of female fertility.

FIG. 7 illustrates a bioinformatics pipeline used to filter through WGS data to identify biomarkers associated with infertility according to certain embodiments. As shown in FIG. 7, samples are subjected to whole genome sequencing, mapping, and assembly. The WGS data is then analyzed to discover genetic variants such as SNPs, small indels, mobile elements, copy number variations, and structural variations. The identified variations are then assessed for statistical significance (See, for example, Tables 2 and 3 above). This includes correction for population stratification, variation-level significance tests, and gene level significance tests. In addition, the biological significance of WGS variants is determined using the SnpEff and Variant Effect Predictor (www.ensembl.org) engines (See, for example, Table 1 above). Variants of biological and statistical significance are then entered into the infertility knowledgebase (i.e., Fertilome database) in order to classify those variants as fertility biomarkers.

The following illustrates use of WGS data to identify variants of interest in accordance with methods of the invention.

Samples were collected from female patients undergoing fertility treatment at an academic reproductive medical center, and categorized into idiopathic infertility or primary ovarian insufficiency (POI) study groups. Phenotypic information was collected for each patient by mining >200 variables from electronic health records. Genomic DNA extracted from blood samples underwent WGS by Complete Genomics (Mountain View, CA). Analysis of genetic variants from WGS was assisted by an infertility knowledgebase with >800 genomic regions of interest (ROI) ranked by a scoring algorithm predicting their likely impact on different fertility disorders, based on publications, data repositories (including protein-protein interactions and tissue expression patterns), meta-analyses of these data, and animal model phenotypes.

The collected female samples were subjected to the processes/algorithms depicted in FIGS. 5-7 (described in more detail above). With those female samples, approximately 50,000 novel variants (approximately 1.6% of total variants observed) were identified as having fertility significances that have not been previously reported in databases such as the sbSNP reference.

5 The identified fertility-related variants included single nucleotide polymorphisms (SNPs, insertions, deletions, copy number variations, inversions, and translocations. Of the SNPs, some of them are predictive to have putative functional significance based on the knowledgebase. For example, the knowledgebase scored some SNPs as deleterious mutations due to potential loss of function or changes in protein structure.

10 In certain aspects, the genomic data, such as WGS data, of a patient/subject population is subjected to a population stratification correction. Population stratification correction accounts for the presence of a systematic difference in allele frequencies between subpopulations in a population possibly due to different ancestry. When conducting population stratification, data is compared to a number (e.g., 1,000) of ethnically diverse individuals as part of the 1000 Genomes
15 Project (100G). Principal components analysis (PCA) is applied to model and identify ancestry differences. In addition, computed association statistics are adjusted for the first two principal components.

FIG. 13 illustrates population stratification correction of two patient groups. The patient groups include female patients undergoing non-donor in vitro fertilization (IVF) cycles. The
20 patients were 38 years old or younger at the time of enrollment, and had no history of carrying a pregnancy beyond the first term before IVF treatment. Each patient had lack of an apparent cause for infertility (i.e., unexplained) after an evaluation of a complete medical history, physical examination, endocrine profile, and the results of an intimate partner's sperm analysis. The patients were divided into two groups. Group A included 11 patients that experienced no live
25 birth or pregnancy beyond the first trimester after 3 or more IVF cycles. Group B included 18 patients that experienced live birth or pregnancy beyond the first trimester through use of IVF therapy. With population stratification correction, Group A and B patients cluster (are shown as black dots) with East Asian, African, Hispanic, and European individuals as shown in the principal component analysis chart of FIG. 13. This data shows that ethnicity may be linked to
30 infertility, or that certain genomic variations are more prevalent in certain ethnic populations. Accordingly, aspects of the invention involve assessing ethnicity of an individual, either through

self-reporting by the individual (e.g., by a questionnaire) or via an assay that looks for known biomarkers related to genetic ethnicity of an individual. That ethnicity data (genetic or self-reported) may be used to guide testing, such as by ensuring that certain genomic variations are checked that are known to be associated with certain ethnic populations.

5

Example 19

Approximately 15% of couples experiencing difficulty conceiving are diagnosed with idiopathic infertility. Genetic polymorphisms could shed light on many of these currently unexplained cases by revealing disruptions to oocyte quality or uterine receptivity that may exist on a subcellular level.

10

In accordance with certain aspects, copy number variations are examined for their effect on female fertility using comparative genomic hybridization (CGH) arrays. CGH provides for methods of determining the relative number of copies of nucleic acid sequences in one or more subject genomes or portions thereof (for example, an infertility marker) as a function of the location of those sequences in a reference genome (for example, a normal human genome). As a result, CGH provides a map of losses and gains in nucleic acid copy number across the entire genome without prior knowledge of specific chromosomal abnormalities. Methods of the invention capitalize on the ability to detect copy number variations without the need for prior knowledge in order to detect potential mutations with infertility significance within patient populations that have unexplained infertility.

15

20

The following illustrates use of CGH arrays to identify copy number variants of interest in accordance with methods of the invention.

The study examined female patients undergoing non-donor in vitro fertilization (IVF) cycles. The patients were 38 years old or younger at the time of enrollment, and had no history of carrying a pregnancy beyond the first term before IVF treatment. Each patient had lack of an apparent cause for infertility (i.e., unexplained) after an evaluation of a complete medical history, physical examination, endocrine profile, and the results of an intimate partner's sperm analysis. The patients were divided into two groups. Group A included 11 patients that experienced no live birth or pregnancy beyond the first trimester after 3 or more IVF cycles. Group B included 18 patients that experienced live birth or pregnancy beyond the first trimester through use of IVF therapy.

25

30

FIG. 9 provides CGH array data of copy number variations detected in the study populations within statistically significant regions associated with infertility (i.e., copy number variations within the Fertilome nucleic acid). FIG. 10 illustrates a specific copy number variation detected in the GJC2 gene of Chromosome 1 within Groups A and B. This region is specifically expressed in both the oocyte and brain, and is known to be associated with embryo issues. As shown, the region within GJC2 showed deletion in the most infertile patients. FIG. 11 illustrates a specific copy number variation detected in the CRTCL and GDF1 genes of Chromosome 19 within Groups A and B. CRTCL is associated with ovary, oocyte, endometrium, and placenta expression. GDF1 is associated with defects in the formation of anterior visceral endoderm and mesoderm. As shown, both patient groups exhibit copy number deletions in those genes. FIG. 12 illustrates a specific copy number variation detected in a non-coding region of Chromosome 6. As shown, both patient groups exhibit copy number duplication that region.

≡

Incorporation by Reference

References and citations to other documents, such as patents, patent applications, patent publications, journals, books, papers, web contents, have been made throughout this disclosure. All such documents are hereby incorporated herein by reference in their entirety for all purposes.

Equivalents

The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The foregoing embodiments are therefore to be considered in all respects illustrative rather than limiting on the invention described herein. Scope of the invention is thus indicated by the appended claims rather than by the foregoing description, and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced therein.

What is claimed is:

1. A method for assessing whether a genetic region is associated with infertility, the method comprising:
 - identifying a genetic region whose function is suspected of being associated with infertility;
 - producing a genetically modified mouse in which the genetic region whose function is suspected of being associated with infertility is altered; and
 - assessing the mouse for presence of an infertility-associated phenotype, wherein the presence of the phenotype is indicative of the genetic region being associated with infertility.
2. The method according to claim 1, wherein the genetic region comprises a gene.
3. The method according to claim 2, wherein identifying comprises:
 - obtaining data on a set of genetic loci, the set comprising genetic loci known to be associated with infertility and genetic loci having no prior association with infertility; and
 - performing a clustering analysis on the data to identify the genetic loci that have no prior association with infertility that cluster with one or more genetic loci known to be associated with infertility, wherein a genetic loci that has no prior association with infertility that clusters with a genetic loci known to be associated with infertility is classified as a being associated with infertility.
4. The method according to claim 1, wherein data is selected from the group consisting of: gene expression data, phenotype, knowledge of gene pathway, and any combination thereof.
5. The method according to claim 1, wherein the method further comprises:
 - administering a therapeutic agent to the mouse; and
 - assessing the effect of the therapeutic agent on the phenotype.

6. The method according to claim 1, wherein presence of the infertility-associated phenotype is used as a factor in ranking the importance of the gene in a database of genetic loci associated with infertility in humans.
7. The method according to claim 6, wherein presence of the phenotype increases the rank of the gene in the database.
8. The method according to claim 6, wherein absence of the phenotype decreases the rank of the gene in the database.
9. The method according to claim 1, wherein the alteration to the genetic region is a mutation.
10. The method according to claim 9, wherein the mutation is selected from the group consisting of: a single nucleotide polymorphism, a deletion, an insertion, a rearrangement, a copy number variation, and a combination thereof.
11. A method for assessing whether a human genetic alteration is associated with an infertility phenotype in a mouse, the method comprising:
 - identifying a human genetic region whose function is known to be associated with human infertility;
 - producing a genetically modified mouse in which the genetic region whose function is associated with human infertility is altered; and
 - assessing the mouse for presence of the infertility phenotype.
12. The method according to claim 11, wherein the genetic region comprises a gene.
13. The method according to claim 11, wherein the method further comprises:
 - administering a therapeutic agent to the mouse; and
 - assessing the effect of the therapeutic agent on the phenotype.

14. The method according to claim 11, wherein presence of the infertility phenotype is used as a factor in ranking an importance of the gene in a database of genetic loci associated with infertility in humans.

15. The method according to claim 14, wherein presence of the phenotype in the mouse increases the rank of the gene in the database.

16. The method according to claim 14, wherein absence of the phenotype in the mouse decreases the rank of the gene in the database.

17. The method according to claim 11, wherein the alteration to the genetic region is a mutation.

18. The method according to claim 17, wherein the mutation is selected from the group consisting of: a single nucleotide polymorphism, a deletion, an insertion, a rearrangement, a copy number variation, and a combination thereof.