The Human Imprintome v1.0: Over 120 Imprinted Genes in the Human Genome Impose a Major Review on Previous Censuses

*Samara C. Silva-Santiago, Elton José R. Vasconcelos, Diana M. Oliveira**
Programa de Pós-Graduação em Biotecnologia – RENORBIO
Universidade Estadual do Ceará -UECE
Fortaleza, Brasil
*diana.magalhaes@uece.br

Ana Carolina L. Pacheco, Mônica F. Silva
Unidade Acadêmica de Ciências Biológicas e Nutrição, Universidade Federal do Piauí/Campus Senador Helvídio Nunes de Barros- UFPI,
Picos, Brasil
carolandim@ufpi.edu.br

Teresa Cristina L. Rocha, Samyra M.V. Brasil
Lab. DNA Forense, Perícia Forense do Ceará - PEFOCE
Secretaria de Segurança Pública e Desenvolvimento Social - SSPDS
Fortaleza, Brasil
teresacristina@pefoce.ce.gov.br

***Abstract*— A relatively small number of human genes are marked with their parental origin and undergo a process termed "genomic imprinting", which, as a field of study, has grown rapidly in the last 20 years, with a growing figure of around 100 imprinted genes known in the mouse and approximately 120 in the human to comprise the updated whole set of human imprinted genes, the imprintome. Human imprintome is here analyzed by means of a reasonable, valid application of the Semantic Web and Linked Data approaches to a few structured datasets in order to provide a comprehensive collection of all known and predicted imprinted genes in the human genome. We have examined, compiled, structured and linked data to use them as a sharing resource for genome and epigenome interrogated studies regarding imprinted genes. Moreover, we offer our datasets of structured, linked data as being the actual research outcome of this human imprintome analysis because as genomics become more and more data intensive, due to huge amounts of biological data, so does our needs for more structured data to be easier mined and shared. The resulting version of the Linked Human Imprintome is a project within Linked Open Data (LOD) initiative (http://lod-cloud.net/) through Data Hub (http://thedatahub.org/en/dataset/a-draft-version-of-the-linked-human-imprintome).**

Keywords- imprinted genes; imprintome; linked data; structured datasets; genomic imprinting

I. INTRODUCTION

A relatively small number of human genes are marked with their parental origin in a way that only a single parental allele is expressed, what is called monoallelic expression [1]. These genes undergo a process termed genomic imprinting [1], which, as a field of study, has grown rapidly in the last 20 years, with a emergent figure of around 100 imprinted genes known in the mouse and approximately 60 in the human [1]. Imprinted genes are often expressed and imprinted in a tissue- and developmental stage-specific manner. They are dependent on the epigenetic machinery for their initial designation of parental identity, as well as establishment and maintenance of their parent-of-origin-specific gene expression [1]. Epigenetic mechanisms, such as the genomic imprinting, play an essential role in higher eukaryotic (like human) gene regulation [2]. The importance of imprinting goes far beyond its critical relevance as a mechanism to balance parental resource allocation, playing important roles in mammalian development and growth [1]. Many of the newly identified imprinted genes lie within genomic regions linked to the development of diabetes, autism, cancer and obesity, besides the well-known genetic syndromes [3]. Although dispersed throughout the genome, imprinted genes are often found in clusters, which are regulated via a central control element, or imprinting control region (ICR). The epigenetic state of an ICR, and in particular the DNA methylation level has been shown to be crucial for maintaining the imprinted state of genes in a cluster [1]. As the control regions display allele specific DNA methylation, they are also referred to as differentially methylated regions (DMRs) [1]. The whole set of human imprinted genes, termed imprintome (as the great territory in which imprinted genes, DMRs and ICRs are the protagonists), is here analyzed in order to provide a comprehensive collection of all known and predicted imprinted genes in the human genome. As such collection, the human imprintome can be further accessed and screened to give a better picture of such a set of genes that range from those with constitutive monoallelic silencing to those, typically more remote from ICRs and DMRs, that display developmentally regulated, tissue-specific or partial monoallelic expression [4].

Since most imprinted genes have been discovered by distinctly direct approaches, the exact, total number of human imprinted genes is not yet known [1, 5-7]. In this regard, we here provide a reasonable, valid application of the Semantic Web [8] and Linked Data [9] approaches to a few structured datasets regarding the human imprintome. Linked Data refers to a practice for publishing and interlinking structured data on the Web [9], through Semantic Web [8] and Open Data principles. We have stored, organized, filtered, and analyzed massive amounts of existing data on human imprinted genes via resource description framework (RDF) and Bio2RDF [10]. The human imprintome is here presented, therefore, as a large-scale draft map of linked data for enabling the tasks of browsing and mining imprinted genes towards further, domain-expert understanding of their complex nature. We offer our datasets of structured, linked data as being the actual research outcome of this human imprintome analysis because, as genomics become more and more data intensive due to huge amounts of biological data, so does our needs for more structured data to be easier mined and shared among biomedical community. We present the resulting first version of the Linked Human Imprintome as a project within Linked Open Data (LOD) initiative (http://lod-cloud.net/) through Data Hub (http://thedatahub.org/en/dataset/a-draft-version-of-the-linked-human-imprintome).

II. RESULTS AND DISCUSSION

A. The (up-to-date) Whole Human Imprintome Composed of 120 Confirmed, plus 128 Candidate Imprinted Genes

We are aware that, for several features and complications that comprise detecting imprinting status, the actual number of imprinted genes in the human genome remains a matter of debate [1, 7, 11]. Therefore, our report on the *in silico* panorama of the human imprintome comprising 248 (two hundred and forty-eight) human genes, divided in 120 (one hundred and twenty) confirmed by experimental evidences, as shown on Table 1 and Fig. 1, plus 128 candidates (Table 2), is, indeed, a major upgrade to the current estimative [1]. Our compiled results on these 120 human imprinted genes impose a considerable upgrade (around 40% up) of the currently known number (around 60), as mentioned by several reports [1, 12]. Our analyses are neither comprehensive nor complete, but they already encompass newly reported imprinted genes. We, thus, present the distribution of the human imprintome along the chromosomes (Fig. 1), with the interesting feature that, out of these 120 imprinted genes, 30 of them (over 24%) contain ncRNAs, while 16 of them are long ncRNA and 14 are small or micro RNAs. We must recall that 82% of these gene arrangements are conserved in murine. Similar to what has been observed in other recent studies with mice [13], the numbers represent a significant enrichment among imprinted genes for co-regulation by some kind of ncRNA. We have analyzed the human imprintome content on RNA genes because, compared to imprinted protein-coding genes, imprinted ncRNAs show different imprinting features and are more responsible than imprinted protein-coding genes for the mechanism of genomic imprinting. It is imprinted ncRNAs, rather than protein-coding genes, that coexist with large imprinted regions and may contribute to the evolution and regulation of genomic imprinting [13].

B. Influence of CpG, ICR, DMR Sites and Regions on the Human Imprintome

One important component of epigenetic gene regulation is CpG islands (CGIs), which are 500–2000 bp long and associated with promoters in most mammalian genes. Most CpGs in promoters are protected from methylation in somatic tissues [2]. In general, imprinted genes are typically associated with differentially methylated regions (DMRs) that have a constitutional histone signature on the DNA methylated allele and exhibit ~50% DNA methylation, whereby one of the two alleles is DNA methylated depending on the parent of origin [1]. ICRs are associated with blocks of tandem repeats and differential methylation [14] that correspond to these DMRs. Identifying imprinted DMRs in humans is complicated by species- and tissue-specific differences in imprinting status and the presence of multiple regulatory regions associated with a particular gene, only some of which may be imprinted. Other DMR features such as Polycomb complex binding targets and histone markers can also be used [15]. DMRs are distributed across the whole genome, with chromosome 7 containing the highest number - nine DMRs, while chromosomes 13, 21 and Y were the only chromosomes for which no DMRs were identified [15], probably due to the fact that CpG sites were not found following rigid criteria [15]. Conversely, in this regard, we have found a great number of CGIs in both human chromosomes 13 and 21, respectively 5,142 and 2,999 (data not shown). Moreover, the human and mouse genomes contain about the same number of CGIs [16], whereas, for most imprinted DMRs analyzed in humans, the size of the CGIs comprising the DMRs are seemingly all larger in humans than in mouse [16], what corroborates with our own *in silico* analyses performed individually for some newly predicted-to-be imprinted genes, such as sfmbt2 (Scm-like with four motif domain) [17]. Sfmbt2 gene is reported here as a candidate imprinted human gene (Table 2) due to the fact that its CGI content is 2.12 fold larger than its murine counterpart, what can be used as a reasonable feature for prediction as an imprinted candidate in humans.

C. Human Imprintome Content in Protein-Coding and RNA Genes

In a biotype list of these 120 human imprinted genes and 128 candidates, there are 90 protein-coding genes (75%), including 06 retrotransposons, and 30 RNA genes within the 120 genes (data not shown). Interestingly, regarding the 128 imprinted candidates, we can observe that 123 of them (96%) are protein-coding genes, while 4 are non-coding RNAs and 1 is pseudogene (data not shown). This illustrates a likely overestimation for protein-coding genes in bioinformatics predictions of imprinted genes. We have analyzed RNA genes towards their complete characterization within the human imprintome because, since long non-coding (lnc)RNAs can mediate epigenetic silencing of a chromosomal domain in trans, several important implications have arisen for dissecting a few of their functional roles [18]. It is now accepted that macro (or large) ncRNAs are major elements that do have an active role in genomic imprinting and on silencing imprinted genes [13, 19], but their mechanism of action is still poorly understood [19]. Imprinted ncRNAs lie at a crossroad between imprinted genes and ncRNA genes, with common and distinct features characteristic of both. The lncRNAs are a very abundant class of ncRNAs; they are mRNA-like transcripts ranging in length from 200nt to ~100kb lacking significant ORFs [20].

Overall, imprinted genes tend to occur in clusters, and microRNAs are associated with the majority of well-defined clusters of imprinted genes [11, 21]. We show here that there are at least 06 (six) microRNAS among the confirmed 120 imprinted genes, two of them, miR-296 and miR-298, which are also part of the imprinted Gnas/GNAS clusters in both mice and humans. Another member of this growing list of imprinted repeated small RNA gene clusters is C19MC [11], also listed here as part of the human imprintome (Table 1).

Table 1. List of 120 (one-hundred and twenty) human imprinted genes, experimentally confirmed as so by *in vitro* or *ex vivo* studies [references are given either as numbers for regular literature or, for web references, WR1- www.geneimprint.com/; WR2- http://igc.otago.ac.nz/home.html]. Human genes are here distinguished as **Imprinted**, only if they are already recognized as so (either a 'Known Protein coding', a 'Novel Protein coding', a 'Known RNA coding' or a 'Novel RNA coding' genes and are shown in the same way identical to known cDNAs or proteins having an entry in human specific model databases: EntrezGene and HGNC/Vega). The table contains information on HGNC ID/Gene Symbol, Full Gene Name/Description, Chromosome Location, Maternal (M) or Paternal (P) Expression.

HGNC (GENE)	GENE DESCRIPTION	LOCATION	EXP	REFERENCE
12003 (TP73)	Tumor Protein p73	1p36.3	M	[WR1]
687 (DIRAS3)	DIRAS Family, GTP-Binding RAS-Like 3	1p31	P	[WR1]
19408(LRRTM1)	Leucine Rich Repeat Transmembrane Neuronal1	2p12	P	[WR2]
9882(RASSF1)	Ras Association (RalGDS/AF-6) Domain Fam.	3p21.31	M	[54]
3668(FGF12)	Fibroblast Growth Factor 12	3q28	P	[34]

6936(MCCC1)	Methylcrotonoyl-CoA Carboxylase 1 (alpha)	3q27	M/P	[34]
19968(NAP1L5)	Nucleosome Assembly Protein 1-like 5	4q22.1	P	[WR1]
4066(GAB1)	GRB2-Associated Binding Protein 1	4q31.21	P	[29]
10777(SFRP2)	Secreted Frizzled-Related Protein 2	4q31.3	P	[54]
18757(RHOBTB3)	Rho-related BTB domain containing 3	5q15	P	[34]
583(APC)	Adenomatous Polyposis Coli	5q21-q22	P	[34]
11140(SNCB)	Synuclein, Beta	5q35	P	[34]
9370(PRIM2)	Primase, DNA, Polypeptide 2 (58kDa)	6p12-p11.1	M	[WR2]
FAM50B-AS	Family with Sequence Similarity 50, Member B AS	6p25	P	[55]
18789(FAM50B)	Family with Sequence Similarity 50, Member B	6p25.2	P	[34]
9046(PLAGL1)	Pleiomorphic Adenoma Gene-like 1	6q24-q25	P	[WR1]
20956(PHACTR2)	Phosphatase and Actin Regulator 2	6q24.2	M	[21]
5326(HYMAI)	Hydatidiform Mole Associated and Imprinted	6q24.2	P	[WR1]
34515(AIRN)	Antisense of IGF2R RNA (Non-Protein Coding)	6q26	P	[40]
5467(IGF2R)	Insulin-Like Growth Factor 2 Receptor	6q26	M	[40]
10966(SLC22A2*)	Solute Carrier Family 22 (Org. Cat. Trans.)	6q26	M	[WR1]
10967(SLC22A3*)	Solute Carrier Fam. 22 (Ext. Monoamine Trans)	6q26-q27	M	[WR1]
2719(DDC)	Dopa Decarboxylase (Aromatic L-AA Decarb.)	7p12.2	ID	[WR1]
4564(GRB10)	Growth Factor Receptor-Bound Protein 10	7p12-p11.2	ID	[WR1]
11761(TFPI2)	Tissue Factor Pathway Inhibitor 2	7q22	M	[WR1]
10808(SGCE)	Sarcoglycan, Epsilon	7q21-q22	P	[WR1]
14005(PEG10)	Paternally Expressed 10	7q21	P	[WR1]
1440(CALCR)	Calcitonin Receptor	7q21.3	M	[WR2]
14946(PPP1R9A)	Protein Phosphatase 1, Regulatory Subunit 9A	7q21.3	M	[WR1]
2918(DLX5)	Distal-Less Homeobox 5	7q22	M	[WR1]
6553(LEP)	Leptin	7q31.3	M	[34]
15740(CPA4)	Carboxypeptidase A4	7q32	M	[WR1]
7028(MEST)	Mesoderm Specific Transcript Homolog (mouse)	7q32	P	[WR1]
17991(MESTIT1)	MEST Intronic Transcript 1, Antisense RNA	7q32	P	[WR1]
2238(COPG2IT1)	COPG2 Imprinted Transcript 1	7q32	P	[WR1]
2237(COPG2)	Coatomer Protein Complex, Subunit Gamma 2	7q32	M	[34]
23025(KLF14)	Kruppel-Like Factor 14	7q32.3	M	[WR1]
2906(DLGAP2)	Discs, Large (Dros) Homolog-Associated Protein 2	8p23	P	[WR1]
32310(RBM12B)	RNA binding motif protein 12B	8q22		
6283(KCNK9)	Potassium Channel, Subfamily K, Member 9	8q24.3	M	[WR1]
34341(CDKN2B-AS)	CDKN2B Antisense RNA (Non-Protein Coding)	9p21.3		
29(ABCA1)	ATP-Binding Cassette, Sub-Family A (ABC1), M1	9q31.1		[WR1]
30964(ARMC3)	Armadillo Repeat Containing 3	10p12.31	P	[34]
21411(AIFM2)	Apoptosis-Induc Factor, Mitochondrion-Ass 2	10q22.1	P	[34]
23508(STOX1)	Storkhead Box 1	10q22.1	P	[56]
17054(INPP5F)	Inositol Polyphosphate-5-Phosphatase F	10q26.11	P	[WR1]
6295(KCNQ1OT1)	KCNQ1 Opposite Strand/Antisense Transcript 1	11p15	P	[WR1]
10965(SLC22A18AS)	Solute Carrier Fam22(Org. Cation Trans.)M. 18 AS	11p15.5		[53]
33351(miR675	MicroRNA 675	11p15.5	M	[WR2]
4713(H19)	H19, Imprinted Maternally Expressed Transcript	11p15.5	M	[WR1]
5466(IGF2)	Insulin-Like Growth Factor 2 (Somatomedin A)	11p15.5	P	[WR1]
14062(IGF2AS	Insulin-Like Growth Factor 2 AS (Non-Protein Cod)	11p15.5	P	[WR1]
6081(INS)	Insulin	11p15.5	P	[WR1]
6294(KCNQ1)	Potassium Volt-Gat Channel KQT-Like Subfm M1	11p15.5	M	[WR1]
6122(IRF7)	Interferon Regulatory Factor 7	11p15.5		[34]
13335(KCNQ1DN)	KCNQ1 Downstream Neighbor	11p15.4	M	[WR1]
1786(CDKN1C)	Cyclin-Dependent kinase Inhibitor 1C (p57, Kip2)	11p15.5	M	[WR1]
10964(SLC22A18)	Solute Carrier Family 22, Member 18	11p15.5	M	[WR1]
12385(PHLDA2)	Pleckstrin Homology-Like Domain, Fam. A, M2	11p15.5	M	[WR1]
16392(OSBPL5)	Oxysterol Binding Protein-Like 5	11p15.4	M	[WR1]
12796(WT1)	Wilms Tumor 1	11p13	P	[WR1]
21625(ANO1)	Channel	11q13.3	M	[29]
10683(SDHD)	Succinate Dehydrogenase Complex, Subunit D,	11q23	M	[WR2; 30]
10869(ST8SIA1)	Sialyltransferase 1	12p12.1-p11.2	P	[34]
15847(RBP5)	Retinol Binding Protein 5, Cellular	12p13.31	M	[WR1]
33510(HOTAIR)	HOX Transcript Antisense RNA	12q13.3		[45, 57]
18081(WIF1)	WNT Inhibitory Factor 1	12q14.3	P	[54]
9884(RB1)	Retinoblastoma 1	13q14.2	M	[22, 59]
14575(MEG3)	Maternally Expressed 3 (Non-Protein Coding)	14q32	M	[WR1]
2907(DLK1)	Delta-Like 1 Homolog (Drosophila)	14q32	P	[WR1]
32027(MIR431)	microRNA 431	14q32.2	M	[60]
14665(RTL1)	Retrotransposon-like 1	14q32.31	P	[32; 34]
31519(miR134)	microRNA 134	14q32.31	M	[WR2]
32781(SNORD116@)	Small nucleolar RNA, C/D BOX 116 CLUSTER	15q11.2	P	[WR1]
1190(C15orf2)	chromosome 15 open reading frame 2	15q11-q13	NK	[WR2]
7675(NDN)	Necdin Homolog (Mouse)	15q11.2-q12	P	[WR1]
11171(SNURF)	SNRPN Upstream Reading Frame	15q12	P	[WR1]
32771(SNORD107)	Small Nucleolar RNA, C/D Box 107	15q11.2	P	[WR1]
32725(SNORD64)	Small Nucleolar RNA, C/D Box 64	15q12	P	[WR1]
32772(SNORD108)	Small Nucleolar RNA, C/D Box 108	15q11.2	P	[WR1]
32774(SNORD109B)	Small Nucleolar RNA, C/D Box 109B	15q11.2	P	[WR1]
37462(SNHG14)	Small Nucleolar RNA Host Gene 14	15q11.2		
7114(MKRN3)	Makorin Ring Finger Protein 3	15q11-q13	P	[WR1]
4079(GABRA5)	Gamma-Aminobutyric Acid (GABA) A Receptor,	15q11-q12	P	
6814(MAGEL2)	MAGE-Like 2	15q11-q12	P	[WR1]
11164(SNRPN)	Small Nuclear Ribonucleoprotein Polypeptide N	15q11.2	P	[WR1]
32773(SNORD109A)	Small Nucleolar RNA, C/D Box 109A	15q11.2	P	[WR1]
32780(SNORD115@)	Small Nucleolar RNA, C/D Box 115 cluster	15q11.2	P	[WR1]
33343(SNORD115-48)	Small Nucleolar RNA, C/D box 115-48	15q11.2	P	[WR1]
12496(UBE3A)	Ubiquitin Potein Ligase E3A	15q11-q13	M	[WR1]
13542(ATP10A)	ATPase, Class V, Type 10A	15q11.2	M	[WR1]
9875(RASGRF1)	Ras Protein-Specific Guanine Nucleot-Rel Factor	15q24	P	[34]
5465(IGF1R)	Insulin-Like Growth Factor 1 Receptor	15q26.3	NK	[61]
26573(ZNF597)	Zinc Finger Protein 597	16p13	M	[34]
25875(NAA60)	N(alpha)-acetyltransferase 60, NatF catalytic	16p13.3	M	[22]
21316(ANKRD11)	Ankyrin Repeat Domain 11	16q24.3	M	[WR1]
24617(TCEB3C)	Transcription Elongation Factor B Polypeptide 3C	18q21.1	M	[WR1]
31785(C19MC)	Cluster de Micro RNA Cromossomo 19	19		[26]
2976(DNMT1)	DNA (cytosine-5-)-Methyltransferase 1	19p13.2	P	[34]
905(AXL)	AXL Receptor Tyrosine Kinase	19q13.1	M	[62]
12875(ZIM2)	Zinc Finger, Imprinted 2	19q13.4	P	[WR1]
8826(PEG3)	Paternally Expressed 3	19q13.4	P	[WR1]
33464(MIMT1)	MER1 Repeat Containing Imprinted Transcript 1	19q13.4		[32]
13057(ZNF264)	Zinc Finger Protein 264	19q13.4	M	[WR1]
15489(ZNF331)	Zinc Finger Protein 331	19q13.42	M	[21]
22948(NLRP2)	NLR Family, Pyrin Domain Containing 2	19q13.42	M	[WR2]
1055(BLCAP)	Bladder Cancer Associated Protein	20q11.2-q12	ID	[WR1]
7860(NNAT)	Neuronatin	20q11.2-q12	P	[WR1]
15905(L3MBTL)	L(3)MBT-Like 1 (Drosophila)	20q13.12	P	[WR1]
24872(SANG)	GNAS ANTISENSE RNA TRANSCRIPT	20q13.32	P	
24872(GNAS-AS1)	GNAS Antisense RNA 1 (Non-Protein Coding)	20q13.32	P	[WR1]
33634(MIR298)	microRNA 298	20q13.32	P	[46]
31617(MIR296)	microRNA 296	20q13.32	P	[46]
4392(GNASXL)	GNAS Complex Locus	20q13.2-q13.3	P	[58]
4392(GNAS)	GNAS Complex Locus	20q13.3	P	[WR1]
4392(NESP55)	Neuroendocrine Secretory Protein 55	20q13.3	P	[7]
7201(MOV10L1)	Moloney Leukemia Virus 10-Like 1,	22q13.33	M	[34]
12810(XIST)	X (inactive)-specific transcript	Xq13.2	P	[32, 43]
674(ARHGAP4)	Rho GTPase Activating Protein 4	Xq28	M	[34]
23357(MCTS1)	Malignant T Cell Amplified Sequence 1	Xq24	P	[63]

Table 2. List of 128 (one-hundred and twenty-eight) human genes predicted-to-be imprinted (or imprinted candidates), here distinguished as **Predicted**, if they are 'Putative' or 'Candidate" to genomic imprinting based on various methods (comparative and *ab initio* predictions, experimental conflicting evidences, all for which at least one is supported by biological data). Imprinted candidates are shown, having an entry in human specific model databases: EntrezGene and HGNC/Vega. For each candidate, respective references are given either as numbers for regular literature or, for web references, as WR1- www.geneimprint.com/; and WR2 - http://igc.otago.ac.nz/home.html)]. The table contains information on HGNC ID/Gene Symbol, Full Gene Name/Description, Chromosome Location, Maternal (M) or Paternal (P) Expression.

HGNC (GENE)	GENE DESCRIPTION	LOCATION	EXP	REFERENCE
32308 (FAM132A)	Family with Sequence Similarity 132, Member A	1p36.33	M	[34]
3084 (DVL1)	Dishevelled, DSH Homolog 1 (Drosophila)	1p36	M	[WR1]
27916(TMEM52)	Transmembrane Protein 52	1p36.33	P	[WR1]
8851(PEX10)	Peroxisomal Biogenesis Factor 10	1p36.32	M	[WR1]
14000(PRDM16)	PR Domain Containing 16	1p36.23-p33	P	[WR1]
12759(WRAP73)	WD repeat containing, antisense to TP73	1p36.3	M	[WR1]
10315(RPL22)	Ribosomal Protein L22	1p36.3-p36.2	P	[WR1]
4006(FUCA1)	Fucosidase, Alpha-L-1, Tissue	1p34	P	[WR1]
29040(SZT2)	Homo sapiens seizure threshold 2 homolog(mouse)(SZT2)mRNA	1p34.2	M	[34]
1075(BMP8B)	Bone Morphogenetic Protein 8B	1p35-p32	P	[WR1]
4237(GFI1)	Growth Factor Independent 1 Transcription Repressor	1p22	P	[WR1]
33238(CCBL2)	Cysteine Conjugate-Beta Lyase 2	1p22.2	M	[34]
29835(NDUFA4P1)	NADH Dehydrogenase (Ubiquinone) 1 Alpha Subcomplex, 4, 9kDa, Pseudogene 1	1p13.3	P	[WR1]
5239(HSPA6)	Heat Shock 70kDa Protein 6 (HSP70B')	1q23	M	[WR1]
9647(PTPN14)	Protein Tyrosine Phosphatase, Non-Receptor Type 14	1q32.2	M	[WR1]
20514(HIST3H2BB)	Histone Cluster 3, H2BB	1q42.13	M	[WR1]
15719(OBSCN)	Obscurin, Cytoskeletal Calmodulin and Titin-Interacting	1q42.13	P	[WR1]
14998(OR11L1)	Olfactory Receptor, Family 11, Subfamily L, Member 1	1q44	P	[WR1]
2597(CYP1B1)	Cytochrome P450, Family 1, Subfamily B, Polypeptide 1	2p22.2	P	[WR1, 17]
1108(ZFP36L2)	Zinc Finger Protein 36, C3H Type-Like 2	2p22.3-p21	M	[WR1]
13887(ABCG8)	ATP-Binding Cassette, Sub-Family G (WHITE), Member 8	2p21	M	[WR1]
29400(CCDC85A)	Coiled-Coil Domain Containing 85A	2p16.1	P	[WR1]
8521(OTX1)	Orthodenticle Homeobox 1	2p13	M	[WR1]
12661(VAX2)	Ventral Anterior Homeobox 2	2p13	M	[WR1]
4463(GPR1)	Official Symbol: GPR1 and Name:	2q33.3	P	[64]
14523(TIGD1)	Tigger Transposable Element Derived 1	2q37.1	P	[WR1]
21314(MYEOV2)	Myeloma Overexpressed 2	2q37.3	P	[WR1]
3978(ALDH1L1)	Aldehyde Dehydrogenase 1 Family, Member L1	3q21.3	M	[WR1]
12872(ZIC1)	Zic Family Member 1	3q24	M	[WR1]
5192(HES1)	Hairy and Enhancer of Split 1, (Drosophila)	3q28-q29	P	[WR1]
3693(FGFRL1)	Fibroblast Growth Factor Receptor-Like 1	4p16	M	[WR1]
11253(SPON2)	Spondin 2, Extracellular Matrix Protein	4p16.3	P	[WR1]
29304(KIAA1530)	Uncharacterized Protein **KIAA1530**	4p16.3	M	[WR1]
3080(DUX2)	DUX2 double homeobox 2	4q35.2	P	[34]
17108(ADAMTS16)	ADAM Metallopeptidase with Thrombospondin Type 1 Motif	5p15	M	[WR1]
1757(CDH18)	Cadherin 18, Type 2	5p15.2-p15.1	P	[WR1]
24943(FAM174A)	Family with Sequence Similarity 174, Member A	5q21.1	P	[34]
2434 (CSF2)	Colony Stimulating Factor 2 (Granulocyte-Macrophage)	5q31.1	M	[WR1]
1142(BTNL2)	Butyrophilin-Like 2 (MHC Class II Associated)	6p21.3	M	[WR1]
21232(MRAP2)	Melanocortin 2 Receptor Accessory Protein 2	6q14.2	P	[WR1]
21606(BRP44L)	Brain Protein 44-Like	6q27	P	[WR1]
5104(HOXA3)	Homeobox A3	7p15-p14	M	[WR1]
5103(HOXA2)	Homeobox A2	7p15-p14	M	[WR1]
5105(HOXA4)	Homeobox A4	7p15-p14	M	[WR1]
5106(HOXA5)	Homeobox A5	7p15-p14	M	[WR1]
5101(HOXA11)	Homeobox A11	7p15-p14	M	[WR1]
3506(EVX1)	Even-Skipped Homeobox 1	7p15-p14	P	[WR1]
4319(GLI3)	GLI Family Zinc Finger 3	7p13	M	[WR1]
21754(TMEM60)	Transmembrane Protein 60	7q11.23	P	[WR1]
18957(MAGI2)	Membrane Ass. Guanylate Kinase WW and PDZ Domain Cont. 2	7q21	M	[WR1]
31773(MIR335)	microRNA 335	7q32.2	P	[65]
11028(SLC4A2)	Solute Carrier Family 4, Anion Exchanger, Member 2 (Erythrocyte Membrane Protein Band 3-Like 1)	7q35-q36	M	[WR1]
24676(FASTK)	Fas-Activated Serine/Threonine Kinase	7q35	M	[WR1]
17930(PURG)	Purine-Rich Element Binding Protein G	8p11	P	[WR1]
26829(NKAIN3)	Na+/K+ Transporting ATPase Interacting 3	8q12.3	P	[WR1]
13348(LY6D)	Lymphocyte Antigen 6 Complex, Locus D	8q24-qter	P	[WR1]
4552(GPT)	Glutamic-Pyruvate Transaminase (Alanine Aminotransferase)	8q24.3	M	[WR1]
578(APBA1)	Amyloid Beta (A4) Precursor Protein-Binding, Fam A, M1	9q13-q21.1	P	[WR1]
28784(C9orf85)	Chromosome 9 Open Reading Frame 85	9q21.13	P	[WR1]
38603(FAM75D3)	Family with sequence similarity 75, member D3	9q21.32	M	[34]
37283(FAM75D1)	Family with Sequence Similarity 75, Member D1	9q21.32	M	[WR1]
6654(LMX1B)	LIM Homeobox Transcription Factor 1, Beta	9q34	M	[WR1]
28435(C9orf116)	Chromosome 9 Open Reading Frame 116	9q34.3	P	[WR1]
20594(EGFL7)	EGF-Like-Domain, Multiple 7	9q34.3	P	[WR1]
30033(PHPT1)	Phosphohistidine Phosphatase 1	9q34.3	M	[WR1]
20256(SFMBT2)	Scm like with four mbt domains	10p14	P	[66]
4172(GATA3)	GATA Binding Protein 3	10p15	P	[WR1]
6532(LDB1)	LIM Domain Binding 1	10q24-q25	M	[WR1]
27275(C10orf91)	Chromosome 10 Open Reading Frame 91	10q26.3	M	[WR1]
19321(NKX6-2)	NK6 Homeobox 2	10q26	M	[WR1]
25247(TTC40)	tetratricopeptide repeat domain 40	10q26.3	M	[WR1]
13639(VENTX)	VENT Homeobox	10q26.3	M	[WR1]
20837(PAOX)	Polyamine Oxidase (Exo-N4-Amino)	10q26.3	M	[WR1]
5412(IFITM1)	Interferon Induced Transmembrane Protein 1 (9-27)	11p15.5	M	[WR1]
35118(C11orf89)	chromosome 11 open reading frame 89	11p15.5	M	[34]
26315(B4GALNT4)	Beta-1,4-N-Acetyl-Galactosaminyl Transferase 4	11p15.5	M	[WR1]
9025(PKP3)	Plakophilin 3	11p15	M	[WR1]
18370(RAB1B)	RAB1B, Member RAS Oncogene Family	11q12	M	[WR1]
22934(KBTBD3)	Kelch Repeat and BTB (POZ) Domain Containing 3	11q22.3	P	[WR1]
17941(NTM)	Neurotrimin	11q25	P	[WR1]
60(ABCC9)	ATP-Binding Cassette, Sub-Family C (CFTR/MRP), Member 9	12p12.1	M	[WR1]
5130(HOXC9)	Homeobox C9	12q13.3	M	[WR1]
5126(HOXC4)	Homeobox C4	12q13.3	M	[WR1]
14470(SLC26A10)	Solute Carrier Family 26, Member 10	12q13	M	[WR1]
1773(CDK4)	Cyclin-Dependent Kinase 4	12q14	M	[WR1]
23820(E2F7)	E2F Transcription Factor 7	12q21.2	M	[WR1]
29308(FBRSL1)	Fibrosin-Like 1	12q24.33	M	[WR1]
5293(HTR2A)	5-Hydroxytryptamine (Serotonin) Receptor 2A	13q14-q21	P	[WR2]
24754(PRR20A)	Proline Rich 20A	13q21.1	M	[WR1]

28297(FAM70B)	Family with Sequence Similarity 70, Member B	13q34	M	[WR1]
3811(FOXG1)	Forkhead Box G1	14q13	P	[WR1]
15767(FERMT2)	Fermitin Family Member 2	14q22.1	P	[WR1]
14574(MEG8)	maternally expressed 8 (non-protein coding)	14q32.31*	M	
24163(BEGAIN)	brain-enriched guanylate kinase-associated homolog (rat)	14q32.2*	P	[67]
4083(GABRB3)	Gamma-Aminobutyric Acid (GABA) A receptor, Beta 3	15q11.2-q12	P	[68]
6109(IPW)	imprinted in Prader-Willi syndrome (non-protein coding)	15q11-q12	P	
11203(SOX8)	SRY (Sex Determining Region Y)-Box 8	16p13.3	P	[WR1]
10524(SALL1)	Sal-Like 1 (Drosophila)	16q12.1	M	[WR1]
25792(C16orf57)	Chromosome 16 Open Reading Frame 57	16q21	M	[WR1]
25070(ACD)	Adrenocortical Dysplasia Homolog (Mouse)	16q22.1	M	[WR1]
3809(FOXF1)	Forkhead Box F1	16q24	M	[WR1]
32371(TMEM88)	Transmembrane Protein 88	17p13.1	M	[WR1]
9749(PYY2)	Peptide YY, 2 (Seminalplasmin)	17q11	P	[WR1]
5113(HOXB2)	Homeobox B2	17q21-q22	M	[WR1]
5114(HOXB3)	Homeobox B3	17q21.3	M	[WR1]
24265(PTRH2)	peptidyl-tRNA hydrolase 2	17q23.2	M	[34]
26136(FAM59A)	Family with Sequence Similarity 59, Member A	18q12.1	P	[WR1]
14015(CELF4)	CUGBP, Elav-Like Family Member 4	18q12	M	[WR1]
9230(PPAP2C)	Phosphatidic Acid Phosphatase Type 2C	19p13	M	[WR1]
32469(ZNF738)	Zinc Finger Protein 738	19p12	P	[WR1]
30700(TSHZ3)	Teashirt Zinc Finger Homeobox 3	19q12	P	[WR1]
15993(CHST8)	Carbohydrate (N-Acetylgalactosamine 4-0) Sulfotransferase 8	19q13.1	M	[WR1]
13018(ZNF225)	Zinc Finger Protein 225	19q13.2	P	[WR1]
28643(ZNF550)	zinc finger protein 550	19q13.31	M	[34]
13022(ZNF229)	Zinc Finger Protein 229	19q13.31	M	[WR1]
6608(LILRB4)	leukocyte immunoglobulin-Like Receptor, Subfamily B , M 4	19q13.4	M	[WR1]
32166(DPRX)	Divergent-Paired Related Homeobox	19q13.42	----	[21]
30216(CHMP2A)	Charged Multivesicular Body Protein 2A	19q	M	[WR1]
35127(PEG3-AS1)	PEG3 antisense RNA 1 (non-protein coding)	19q13.43	P	[55]
18563(USP29)	ubiquitin specific peptidase 29	19q13.43	----	[69]
16366(ZIM3)	zinc finger, imprinted 3	19q13.4	P	[69]
13108(MZF1)	Myeloid Zinc Finger 1	19q13.4	M	[WR1]
16213(ISM1)	Isthmin 1 Homolog (Zebrafish)	20p12.1	P	[WR1]
16435(HM13)	histocompatibility (minor) 13	20q11.21	----	[70]
15866(C20orf20)	Chromosome 20 Open Reading Frame 20	20q13.33	M	[WR1]
2219(COL9A3)	Collagen, Type IX, Alpha 3	20q13.3	M	[WR1]
10883(SIM2)	Single-Minded Homolog 2 (Drosophila)	21q22.2	P	[WR1]
2846(DGCR6)	DiGeorge Syndrome Critical Region Gene 6	22q11.21	P	[WR1]

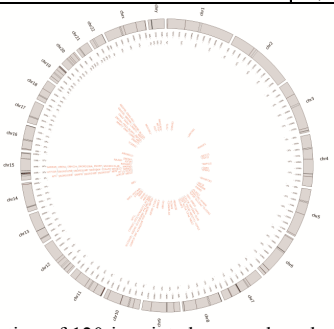

Figure 1. Distribution of 120 imprinted genes along human chromosomes. Each imprinted gene position is represented by thin red lines within chromosomes (gray rectangles), while their symbols are written in red. The white regions displayed correspond to the centromers in each chromosome. This image was generated using the software circus, version 0.55.

III. CONCLUSION

Around 60 imprinted genes were reported earlier in humans (www.geneimprint.com) [1, 5, 6, 15, 22], as opposed to a far greater number of genes that are imprinted in mice (over 100) [1]. Benefiting from previous reports on genomic imprinting data, and on the increasing number of bioinformatics tools developed to explore these data, we have analyzed human imprinted genes data resulting in a first version of the Linked Human Imprintome with a 2x fold increase from that previous number (120 confirmed human imprinted genes, available at http://thedatahub.org/en/dataset/a-draft-version-of-the-linked-human-imprintome). We expect that the Linked Data approach, in concert with recent experimental post-genomics technologies, will lead to a future full characterization of the imprintome in humans.

IV. METHODS

A. *Ethics statement*

Experimental designs, data and procedures were exclusively *in silico* and in agreement with Open Data Commons' guidelines (http://opendatacommons.org/guide/). Pertinent human, mouse and other mammal's biological samples cited (DNA, RNA and protein sequences, structures and interactions) were exactly as provided by their original data sources.

B. *Data Sources, Data Collection and Detailed Annotations*

Exhaustive searches for reviews of imprinted genes and loci were performed using queries from a table of pertinent keywords. A catalog including gene names was compiled and, for each imprinted or predicted-to-be imprinted gene, we extracted annotations, references, and sequences. Imprinting features, such as imprinted region, imprinted tissues and stages, DMRs, ICRs and expressed allele, which could not be found in existing databases, were obtained by analyzing related publications. The compiled annotations were then formalized and imported into a MySQL database with tables for four gene categories: protein-coding or RNA-coding, non-coding (ncRNA) and retrotransposon-coding. Each table was assigned general annotation fields, as well as specific annotation fields for its corresponding category.

A collection of human ICRs/DMRs was extracted from research papers, while datasets of imprinted genes were downloaded from all available repositories such as: i) Geneimprint (www.geneimprint.com/); ii) the Catalog of Imprinted Genes [6]; iii) WAMIDEX [23] (https://atlas.genetics.kcl.ac.uk); iv) ncRNAimprint (http://rnaqueen.sysu.edu.cn/ncRNAimprint); v) HGNC (HUGO Gene Nomenclature Committee) (www.genenames.org); besides vi) patents which involve imprinted genes [5, 24]. Additional data about imprinted genes were collected, using gene names or aliases as queries, from Online Mendelian Inheritance in Man (OMIM) (www.ncbi.nlm.nih.gov/omim/), Genome-Wide Association Studies (GWAS) Catalog (www.genome.gov/gwastudies), GeneCards® (www.genecards.org); and Diseasome linked data (www4.wiwiss.fu-berlin.de/diseasome/). All data were collected in annotation-rich genomic context and/or in gene-centric view to benefit from many relevant annotations and discriminations among imprinted, predicted-to-be imprinted and not-imprinted genes. Both validated and predicted data on imprinted genes were collected and imported into our datasets. Sequences were downloaded from the human (GRCh37.p5/hg19) assembly using UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

C. *Data Management Operations*

Human Vega 45 (http://vega.sanger.ac.uk/) contains an update to manual annotation of the human genome and gene structures are presented in the merged human geneset shown in Ensemble (corresponding to Gencode - www.gencodegenes.org/). All annotation is from the Havana group (Human and Vertebrate Analysis and Annotation) at the Wellcome Trust Sanger Institute, including 'Sanger chromosomes', i.e., those originally sequenced at the Institute (chromosomes 1,6,9,10,11,13,20,22) and target regions of ENCODE pilot project (www.genome.gov/26525202). For visualization of genomic data, specifically for detailed features on gene names, genes involved in disease, cancer mutations and variable regions, Artemis Genome Browser was used (http://sanger.ac.uk/resources/software/artemis/). Before adopting the model defended by LOD cloud diagram and projects (http://lod-cloud.net/), via a SPARQL endpoint (approximately 130,000 triples in a N3 file format), S3QL (http://s3ql.info) and S3DB data service (http://link.s3db.org/owl) were used.

D. *In Silico Prediction of Imprinted Genes*

As described [5, 22], we have extended the search to epigenomic features [22], using high-confidence predictions [5] and also ncRNA genes [13]. We have applied the local CpG density, CpGi130, [25], with a window of 500 bp. As described

[25], we used a weighted count of CpG sites in the genome upstream and downstream 500 bases from a given point of interest (e.g., nearby imprinted loci). Weight decreases linearly from 1 at the center of the point of interest to 0 at 500 bases up- or downstream. The score reveals the number of CpG sites in close proximity to the point of interest.

E. Data Release

The first version of the Linked Human Imprintome, a project within LOD initiative through Data Hub (http://thedatahub.org/en/dataset/a-draft-version-of-the-linked-human-imprintome), is available under PDDL v1.0 (Public Domain Dedication and License (opendatacommons.org/licenses/pddl/1.0/). Although PDDL licenses tend to mention "Databases" rather than "Data", they are entirely suited to "data" or "datasets" and, in juridical terminology, the preferred term for a collection of information is "Database" rather than "data".

REFERENCES

[1] J. M. Frost, and G. E. Moore, "The importance of imprinting in the human placenta," PLoS Genet., vol. 6, Jul. 2010.

[2] E. Schneider et al., "Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns," Nucleic Acids Res., vol. 12, Jul. 2010, pp. 3880-3890.

[3] I. Lobo, "Genomic imprinting and patterns of disease inheritance," Nature Educ. Vol. 1, 2008.

[4] E. Ivanova, and G. Kelsey, "Imprinted genes and hypothalamic function," J. Mol. Endocrinol., vol. 47, Sep. 2011, pp. R67-R74.

[5] R. L. Jirtle, A. J. Hartemink, and P. P. Luedi, "Patent Application Publication: Imprinted genes and diseases," US2011/0014607A1, Jan. 2011, pp. 1-13.

[6] I. M. Morison, C. J. Paton, and S. D. Cleverley, "The imprinted gene and parent-of-origin effect database," Nucleic Acids Res., vol. 1, Jan. 2001, pp. 275-276.

[7] K. Nakabayashi et al., "Methylation screening of reciprocal genome-wide UPDs identifies novel human-specific imprinted genes," Hum. Mol. Genet., vol. 20, Aug. 2011, pp. 3188-3197.

[8] T. Berners-Lee, J. Hendler, and O. Lassilia, "The semantic web," Scientific American, vol. 5, May 2001, pp. 34-44.

[9] T. Heath, and C. Bizer, Linked data: evolving the web into a global data space, 1st ed., vol. 1. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011, pp. 1-136.

[10] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault,and J. Morisset, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems," J. Biomedical Informatics, vol. 41**,** Mar. 2008, pp. 706-716.

[11] M. Noguer-Dance et al., "The primate-specific microRNA gene cluster (C19MC) is imprinted in the placenta," Hum. Mol. Genet.,vol. 19, Sep. 2010, pp. 3566-3582.

[12] D. Monk, "Deciphering the cancer imprintome," Brief. Funct. Genomics, vol. 9, Jul. 2010, pp. 329-339.

[13] Y. Zhang et al., "ncRNAimprint: A comprehensive database of mammalian imprinted noncoding RNAs," RNA*,* vol. 16, Oct. 2010, pp. 1889-1901.

[14] B. Hutter, V. Helms, and M. Paulsen, "Tandem repeats in the CpG islands of imprinted genes," Genomics*,* vol. 88, Sep. 2006, pp. 323-332.

[15] R. K. C. Yuen, R. Jiang, M. S. Peñaherrera, D. E. McFadden, and W. P. Robinson, "Genome-wide mapping of imprinted differentially methylated regions by DNA methylation profiling of human placentas from triploidies," Epigenetics Chromatin, vol. 4, Jul. 2011, pp. 10.

[16] R. S. Illingworth et al., "Orphan CpG islands identify numerous conserved promoters in the mammalian genome," PLoS Genet., vol. 6, Sep. 2010.

[17] A. Kuzmin et al., "The PcG gene Sfmbt2 is paternally expressed in extraembryonic tissues," Gene Expr. Patterns, vol. 8, Jan. 2008, pp. 107-116.

[18] R. Huang et al., "An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs," PLoS One., vol. 11, Nov. 2011.

[19] C. M. Williamson et al., "Uncoupling antisense-mediated silencing and DNA methylation in the imprinted GNAS cluster," PLoS Genet., vol. 3, Mar. 2011.

[20] E. A. Gibb, C. J. Brown, and W. L. Lam, "The functional role of long non-coding RNA in human carcinomas," Mol. Cancer, vol. 10, Apr. 2011, pp. 38-54.

[21] J. E. Robson, S. A. Eaton, P. Underhill, D. Williams, and J. Peters, "MicroRNAs 296 and 298 are imprinted and part of the GNAS/Gnas cluster and miR-296 targets IKBKE and Tmed9," RNA, vol. 18, Jan. 2012, pp. 135-144.

[22] C. M. Brideau, K. E. Eilertson, J. A. Hagarman, C. D. Bustamante, and P. D. Soloway, "Successful computational prediction of novel imprinted genes from epigenomic features," Mol. Cell Biol., vol. 30, Jul. 2010, pp. 3357-3370.

[23] R. Schulz et al., "WAMIDEX: a web atlas of murine genomic imprinting and differential expression," Epigenetics, vol. 3**,** Mar. - Apr. 2008, pp. 89-96.

[24] A. P. Feinberg, L. Strichman-Almashanu, and S. Jiang, "Patent: Gene Imprinting and Methylated CpG islands," US7488815B2, Feb. 2009, pp. 1-41.

[25] M. D. Robinson et al., "Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation," Genome Res., vol. 20, Dec. 2010, pp. 1719-1729.