2010 March 14th Nature submitted

Theoetical analysis indicates human genome is not a blueprint and human oocytes have the instructions.

Koichi Itoh

The Institute for Theoretical Molecular Biology

21-13, Rokurokuso-cho, Ashiya, Hyogo, JAPAN 659-0011

TEL: +81-797-35-6368 FAX: +81-797-35-6368

http://www.i-tmb.com/

e-mail: koichiitoh@yahoo.co.jp

Corresponding author: Koichi Itoh

The Institute for Theoretical Molecular Biology

21-13, Rokurokuso-cho, Ashiya, Hyogo, JAPAN 659-0011

TEL: +81-797-35-6368

FAX: +81-797-35-6368

e-mail: koichiitoh@yahoo.co.jp

http://www.i-tmb.com/

Human genome has been thought to be a blueprint, but what type of the blueprint has been a mystery. Human genome project was over in 2003, and seven years are already passed, but the number of human genes still unknown. Analysis of human genomes has been continuously done, but the discussion which a human genome is a blueprint has not been done. Far from that, any traces of a blueprint are not found in human genomes. This may be evidence that a human genome is not a blueprint. The Watson-Click's DNA double helix is very beautiful. Hence, we life-scientists have been imprinted that a human genome is a blueprint. If we hypothesize that a human genome is a blueprint, what types of absurdity do emerge? And if a human genome is not a blueprint, what must be needed to construct human bodies? To solve these hypothetical propositions are the aim of this document. In the case of unicellular organisms such as *E.coli*, their genomes may play a role for blueprints. However, biological mechanisms of multicellular organisms such as *Homo Sapiens*, are much complex and it is difficult to contain all information as a blueprint in their genomes. Therefore, a human genome plays a role for storage of genes, and I think that human oocytes have the instructions and a fertilized egg selects necessary genes from that storage, and expresses genes for development and differentiation.

Human genome is not a blueprint. At first the definition of a blueprint must be determined. According to a dictionary, a blueprint for something is a plan or set of proposals that shows how it is expected to work. I scrutinized loci of genes for 8 important biological pathways and factors, and their loci are scattered all over the human genome at random. (Table I) I think that a blueprint must have regularity, periodism, harmony, some types of patterns, consistency or beauty which a blueprint itself has. But there were not existed such things. On the contrary, more than half of human genome sequence consists of Lines, Sines, retroviral-like elements, DNA-only transposon fossils, *Alu* sequences and pseudogenes[1.] The loci of genes for 8 pathways and factors are scattered all over the human genome, and there do not exist any operons such as in bacterial genomes. Some reports exist that genes that make a cluster in one-dimensional, construct a cluster in three-dimensional, but there are no report that scattered genes in one-dimensional construct a cluster in three-dimensional[2]. In mathematics, one opposite example is enough for proof. But biology has some exceptions. However, genes in Table I are biologically important genes, and if a human genome is a blueprint, 8 exceptions must not be permitted. Here, I logically show that a human genome is not a blueprint. Hence, how are human bodies constructed from a human genome which is storage of genes?

Human oocytes have the instructions. Before fertilization, human oocytes express genes. If a human genome is storage of genes, mRNAs which are important for development and differentiation must be expressed in human oocytes and translated into proteins as soon as fertilization begins. Therefore, I surveyed public databases and I found an expression profile in human oocytes. In that profile, there are 12700 genes, and among 12700 genes, I found more than 800 genes which are related to development and differentiation. In general, many sample data must be necessary for comparison of gene expression levels for statistical analysis. But in my case, I do not need statistical analysis. Because the importance is only in which certain types of genes are expressed in human oocytes. I think that human oocytes play a major role because of the amount of genes related to development and differentiation. Essential genes for human development and differentiation such as *Oct3, Oct4* are not existed in Table II. But I do not think that it is critical. I just think that mRNAs of *Oct3, Oct4* did not hybridize on the microarray chips. Because the genes which must be expressed must be expressed in human oocytes. And because of RNA interference, some mRNA might be broken. However, the amount of genes in human oocytes related in development and differentiation indicates that human oocytes have the instructions. Definition of instruction must be done. Instructions are clear and detailed information on how to do something. In this point, I think that human oocytes have the simple instructions. If human oocytes do not have the simple instructions, where is the blueprint or the instructions? I already indicate that a human genome is not a blueprint. Hence, it is logical that human oocytes have the simple instructions because a human body begins to be built from only one cell, a fertilized egg. If other cells except for human oocytes give proteins or mRNAs from outside of human oocytes, nurse cells or stromal cells might be candidates for the simple instructions. But it is not realistic that those cells give most of biologically important proteins or mRNAs into fertilized eggs. Therefore, I logically proved that human oocytes have the simple instructions.

Important genes for the instruction in human oocytes[3-7]. The homeodomain is an approximately 60 amino acid sequence containing many basic residues, and forms a helix-turn-helix structure that binds specific sites in DNA. The homeodomain sequence itself is coded by a corresponding homeobox (HOX) in the gene. The homeobox was given its name because it was initially discovered in homeotic genes. However, there are many transcription factors that contain a homeodomain as their DNA-binding domain and although they are often involved in development, possession of a homeodomain does not guarantee a role in development, nor are mutants of homeobox genes necessarily homeotic.

A very large number of homeodomain proteins have important functions, e.g. Engrailed in *Drosophila* segmentation, Goosecoid in the vertebrate organizer, Cdx proteins in anteroposterior patterning. An important subset are the HOX proteins which have a special role in the control of anteroposterior pattern in animals. Homeobox genes are found in animals, plants, and fungi, but the Hox subset are only found in animals. The LIM domain is a cysteine-rich zinc-binding region responsible for protein-protein interactions, but is not itself a DNA-binding domain. LIM-homeoproteins possess two LIM domains together with the DNA-binding homeodomain. Examples are Lim-1 in the organizer, Islet-1 in motorneurons, Lhx factors in the limb bud, and Apterous in the *Drosophila* wing. PAXs are characterized by a DNA-binding region called a paired domain with 6 alpha-helical segments. The name is derived from the paired protein in Drosophila. Many of pax proteins also contain a homeodomain. Examples are Pax6 in the eye and Pax3 in the developing somite. Zinc-finger protein is a large and diverse group of proteins in which the DNA-binding region contains projections (“fingers”) with Cys and/or His residues folding around a zinc atom. Some examples are the GATA factors important of the blood and the gut, Krupple in the early *Drosophila* embryo, WT-1 in the kidney. Basic helix-loop-helix (bHLH) protein transcription factors are active as heterodimers. They contain a basic

DNA-binding region and a hydrophobic helix-loop-helix region responsible for protein dimerization. One member of the dimer is found in all tissues of the organism and the other member is tissue specific. There are also proteins containing the HLH but not the basic part of the sequence. These form inactive dimmers with other bHLH proteins and so inhibit their activity. Examples of bHLH proteins include E12, E47 which are ubiquitous in vertebrates, the myogenic factor MyoD, and *Drosophila* pair-rule protein hairy. An inhibitor with no basic region is Id, which is an inhibitor of myogenesis. FOX have a 100 amino acid winged helix domain which forms another type of DNA-binding region and known as "FOX" proteins. Examples are Forkhead in *Drosophila* embryonic termini and Fox2A in the vertebrate main axis and gut. T-box factors have a DNA-binding domain similar to the prototype gene product known as "T" in the mouse and as brachyury in other animals. They include the endodermal VegT and the limb identity factors Tbx4 and Tbx5. High mobility group (HGM)-box factors differ from most others because they do not have a specific activation or repression domain. Instead they work by bending the DNA to bring other regulatory sites into contact with the transcription complex. Examples are SRY, the testis-determining factor, Sox9, a "master switch" for cartilage differentiation, and the TCF and LEF factors whose activity is regulated by the Wnt pathway. Trnasforming growth factor (TGF) beta was originally discovered as a mitogen secreted by “transformed” (cancer-like) cells. It has turned out to be the prototype for a large and diverse superfamily of signaling molecules, all of which share a number of basic structural characteristics. The mature factors are disulfide-bonded dimmers of approximately 25 kDa. They are synthesized as longer pro-forms which need to be protrolytically cleaved to the mature form in order for biological activity to be shown. The TGF-beta themselves are in fact often inhibitory to cell division and promote the secretion of extracellular matrix materials. They are involved mainly in the organogenesis stages of development. The activin-like factors include the nodal-related family, which are all involved in induction and patterning of the mesoderm in vertebrate embryos. The bone morphogenetic proteins (BMPs) were discovered as factors promoting ectopic formation of cartilage and bone in rodents. They are involved in skeletal development, and also in the specification of the early body plan. There are a number of receptors for the TGF-beta superfamily. Their specificity for different factors is complex and overlapping, but in general different subsets of receptors bind to the TGF-beta themselves, the activin-like factors, and the BMPs. In all cases the ligand binds first to a type II receptor and enables it form a complex with a type I receptor. The type I receptor is a Ser-Thr kinase and becomes activated in the ternary complex. Activation causes phosphorylation of smad proteins in the cytoplasm. Smads 1, 5, and 8 are targets for BMP receptors; smad 2 and 3 for activin receptors. Smad 4 is required by both pathways, and smad 6 is inhibitory to both by displacing the binding of smad 4. Phosphorylation causes the smads to migrate to the nucleus where they function as for transcription factors, regulating target genes. The hedgehogs were first identified because mutations of the gene in *Drosophila* disrupted the segmentation pattern and made the larvae look like hedgehogs. Sonic hedgehog is very important for the dorsoventral patterning of the neural tube and for anteroposterior patterning of the limbs. Indian hedgehog is important in skeletal development. The full-length hedgehog polypeptide is an autoprotease, cleaving itself into an active N-terminal and an inactive C-terminal part. The N-terminal fragment is normally modified by covalent addition of a fatty acyl chain and of cholesterol, which are needed for full activity. The hedgehog receptor is called patched, again named after the phenotype of the gene mutation in *Drosophila*. This is of the G-protein-linked class. It is constitutively active and is repressed by ligand binding. When active it represses the activity of another cell membrane protein, smoothened, which in turn represses the proteolytic cleavage of Gli-type transcription factors. Full-length Gli factors are transcriptional activators that can move to the nucleus and turn on target genes, but the constitutive removal of the C-terminal region makes them into repressors. In the absence of hedgehog, patched is active, smoothened inactive, and Gli inactive. In the presence of hedgehog, patched is inhibited, smoothened is active, and Gli is active. Activation of protein kinase A also represses Gli and hence antagonizes hedgehog signaling. The founder member of the Wnt family was discovered through two routes, as an oncogene in mice and as the wingless mutation in *Drosophila*. Wnt factors are single-chain polypeptides containing a covalently linked fatty acyl group which is essential for activity and renders them insoluble in water. The Wnt receptors are called frizzled after another Drosophila mutation. There are several classes of receptor for different ligand types and they do not necessarily cross-react. Wnt 1, 3A, or 8 will activate frizzleds that cause the repression of a kinase, glycogen synthase kinase 3 (gsk3) via multifunctional protein called dishevelled. When active, gsk3 phosphorylates beta-catenin, an important molecule involved both in cell adhesion and gene regulation. When gsk3 is repressed, beta-catenin remains unphosphorylated and in this state can combine with a transcription factor, Tcf-1, and convey it into the nucleus. This pathway is important in numerous developmental contexts, including early dorsoventral patterning in *Xenopus*, segmentation in a *Drosophila*, and kidney development. Other Wnts, including Wnts 4, 5, and 11, bind to a different subset of frizzled that activate two other signal transduction pathways. In the planar cell polarity pathway a domain of the dishevelled protein interacts with small GTPases and the cytoskeleton to bring about a polarization of the cell. In the Wnt-Ca pathway phospholipase C becomes activated by a frizzled. This then acts to generate diacylglycerol and inositol 1,4,5 triphosphate, with consequent elevation of cytoplasmic calcium, as described above under G-protein-coupled receptors. For the Delta-Notch system both the ligand (Delta, Jagged) and receptor (Notch) are integral membrane proteins. Their interaction can therefore only take place if the cells making them are in contact, as for the ephrin-Eph system. Binding of ligand to Notch causes cleavage of the cytoplasmic portion of Notch by an intramembranous protease, gamma-secretase, and this causes release into the cytoplasm of transcription factor, CSL-kappa. This migrates to the nucleus and activates target genes. The gamma-secretase is the same protease that generates the peptide whose accumulation in the brain leads to Alzheimer’s disease. Notch can carry O-linked tetrasaccharides and presence of this carbohydrate chain can affect its specificity, increasing sensitivity to Delta and reducing sensitivity to Jagged. Control is often exercised through the activity of the glycosyl transferase Fringe, which adds GlcNAc to the O-linked fucose. The Delta-Notch system is important in numerous developmental situations, including neurogenesis, somitogenesis, and imaginal disc development. Cadherins are families of single-pass transmembrane glycoproteins which can adhere tightly to similar molecules on other cells in the presence of calcium. Cadherins are the main factors attaching embryonic cells together, which is why embryonic tissues can often be caused to disaggregate simply by removal of calcium. The cytoplasmic tail of cadherins is anchored to actin bundles in the cytoskeleton by a complex including proteins called catenins. One of these, beta-catenin, is also a component of the Wnt signalimg pathway, providing a potential link named for the tissues in which they were originally found, so E-cadherin occurs mainly in epithelia and N-cadherin occurs mainly in neural tissue. The integrins are cell-surface glycoproteins that interact mainly with components of the extracellular matrix. They are heterodimers of alpha- and beta- subunits, and require either magnesium or calcium for binding. There are numerous different alpha and beta chain types and so there is a very large number of potential heterodimers. Integrins are attached by cytoplasmic domains to microfilament bundles, so, like cadherins, they provide a link between the outside world and the cytoskeleton. They are also thought on occasion to be responsible for the activation of signal transduction pathways and new gene transcription following exposure to particular extra cellular components.

After central dogma. After the birth of molecular biology, we life-scientists proved only two things, in my opinion. Firstly, there is high possibility that genes or proteins which have similar nucleic acid or amino acid sequences have similar 3-demensional structures and functions. Secondly, Genes or proteins have many functions because of the timing of working, permutation and combination. The number of human genes might be 40000 at most. In the first place, only 40000 genes cannot control complex biological mechanisms. Therefore, I think that limited number of genes and proteins change the timing of working, permutation and combination, and control the diverse biological mechanisms in human bodies. Genomes of viruses or bacteria might have the possibility that those genomes play a role for blueprints. But it will become impossible that human genome play a role for a blueprint. Hence, I think that human genome begins to exist as storage of genes. And human oocytes express essential genes for development and differentiation as the simple instructions. After fertilization, a fertilized egg differentiates according to micro-environment surround the fertilized egg. Therefore, human oocytes expresses genes for adhesion molecules such as integrins, cadherins and so on. From now on, a lot of evidence will be piled up to support my hypothesis. Finally, I foresee that once organogenesis begins, tissue differentiation proceeds autonomously and human bodies are built. This is, I think, theoretical molecular biology.

Reference

1. Alberts, B. Johnson, A. Walter, P. Lewis, J. Raff. M, Roberts, K. Molecular Biology of the Cell, Fifth Ed., 2008 Garland Science. Mortimer Street, London.
2. Schneider R, Grosschedl R. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. Genes Dev 2007;21:3027-3043.
3. Slack, J.M.W. 2006, Essential Developmental Biology 2nd Ed., Blackwell Publishing. West Sussex, UK.
4. Gilbert, S.F. 2006, Developmental Biology, eighth Ed., Sinauer Association Inc. Sunderland, MA.
5. Schoenwolf, G.C. *et al.* 2009, Larsen’s Human Embryology, Fourth Ed., Churchill Livingstone. New York
6. Wolpert, L. 2007, Principles of Development, Third Ed., Oxford University Press.
7. Moody, S.A. 2007, Principles of Developmental Genetics, Academic Press. New York.
8. Kocabas, A. M., Crosby, J., Ross, P.J. *et al.* 2006, The transcriptome of human oocytes, *Proc. Natl Acad. Sci. USA*,. 103, 14027-14032

Materials and Methods

Table I was made from NCBI database (http://www.ncbi.nlm.nih.gov/) and KEGG (http://www.kegg.jp/ja/). One hundred ninety six key words in Supplemental Table I were selected from reference3-7. Supplemental Table II was made from Supplementary Data 1, 2, 3 which were originally located in http://www.canr.msu.edu/dept/ans/community/people/cibelli_jose.html[8]. I re-locate Supplementary Data 1, 2, 3, in http://www.i-tmb.com/text.html. Supplementary Data 1 contains up-regulated genes in human oocytes, Supplementary Data 2 contains down-regulated genes in human oocytes, and Supplementary Data 3 contains uniquely expressed genes in human oocytes. I combined Supplementary Data 1, 2, 3, and eliminated duplicated genes. Finally, I got 12764 genes which expressed in human oocytes (Supplemental Table II). I surveyed 12764 genes with 196 key words and I selected 823 genes which are thought to be important in development and differentiation in GenBank release 175.0 (Supplemental Table III). Table II shows the number of important genes for development and differentiation.