Structure–Function Correlations of the Insulin-Linked Polymorphic Region*

Cytosine-Rich Strands of the Insulin Minisatellite Adopt Hairpins with Intercalated Cytosine$^+$ · Cytosine Pairs†

*Research by *P. Catasti, X. Chen, R.K. Moyzis, E.M. Bradbury, and G. Gupta,* J. Mol. Biol. ***1996,*** *264, 534; †P. Catasti, X. Chen, L.L. Deaven, R.K. Moyzis, E.M. Bradbury, and G. Gupta,* J. Mol. Biol. ***1997,*** *272, 369*

Condensation and commentary by ***Tomas Simonsson, Hans R. Widlund,*** *and* ***Mikael Kubista,*** *Chalmers University of Technology, Sweden*

Condensation of the Research

Purpose of the Studies

To characterize higher order DNA structures that form in the upstream control region of the insulin gene by high-resolution nuclear magnetic resonance spectroscopy

Researchers' Approach

For the guanine-rich strand of the insulin-linked polymorphic region (ILPR), gel electrophoresis and circular dichroism were used to obtain indirect evidence for formation of guanine-quartet structures. One- and two-dimensional nuclear magnetic resonance (NMR) spectroscopy and molecular modeling were used to reveal the high-resolution structure of an intramolecularly folded DNA structure with a core of stacked guanine-quartets. Analogously, for the complementary cytosine-rich strand of the ILPR, one- and two-dimensional NMR in combination with molecular modeling allowed identification of a structure built on intercalating hemiprotonated cytosine–cytosine base pairs, the *i* motif. Finally, in vitro DNA replication was used to demonstrate that the propensity to form either of the two DNA structures increases with the length of the ILPR.

Background

Human type I diabetes, or insulin dependent diabetes mellitus (IDDM), is a disease characterized by the autoimmune destruction of insulin-secreting β cells of the pancreatic islets of Langerhans. With a peak age-at-onset of 12 years, it is one of the fastest growing childhood

CHEMTRACTS—BIOCHEMISTRY AND MOLECULAR BIOLOGY 12:210–217 (1999)

CCC 1431-9268

diseases and causes a lifelong dependence on external sources of insulin. If a molecular basis for IDDM could be elucidated, the development of functional gene therapy might provide an attractive alternative to daily insulin injections.

The sequence of the insulin gene was revealed in 1980[1] and its cytogenetic location was assigned to chromosome 11p15.5.[2] The upstream control region flanking the insulin gene was soon found to be polymorphic in length and to consist of a variable number of tandem repeats with the sequence $(^{CCCCACACCCCTGT}_{GGGGTGTGGGGACA})_n$.[3] This region has been shown both to confer genetic susceptibility to IDDM[4,5] and to regulate expression of the insulin gene.[6,7] Previous biochemical characterization of the ILPR has indicated guanine-quartet formation,[8–11] but the exact nature of possible ILPR structures remained largely indefinite until the present publications revealed at high resolution how quartet formation occurs.

What Researchers Accomplished

Nondenaturing gel electrophoresis studies suggested that the synthetic 25-base-long oligonucleotide d(GGGGTGTGGGGACAGGGGTGTGGGG) was capable of forming an intramolecularly folded guanine-quartet structure in presence of NaCl. Circular dichroism analysis of the same oligonucleotide in the presence of NaCl resulted in a melting temperature of ~85°C, which was consistent with the high melting temperatures normally observed for guanine-quartet structures.[12] The gross morphology of the folded structure was further investigated by P1 nuclease digestion. It was found that both the ACA loop and the two TGT loops were susceptible to digestion, whereas the presumed guanine-quartet core was resistant to hydrolysis. One- and two-dimensional NMR studies of the d(GGGGTGTGGGGACAGGGGTGTGGGG) oligonucleotide unambiguously identified an intramolecular structure with four planes of stacked guanine-quartets (Fig. 1). While normal B-DNA has one minor and one major groove, guanine-quartet structures have four grooves of which two are narrow and two are wider. Here GGGG stretches producing wider grooves are linked by antiparallel TGT loops that protrude on top of the guanine-quartet core. The ACA loop on the other side connects the central two GGGG stretches to yield one of the narrow grooves. The glycosidic torsion angles of the guanines in each GGGG stretch were found to adopt alternating *syn-anti* conformations while all loop-nucleotides were *anti*. The sugar puckers for all nucleotides in the folded structure were C2′-endo.

In a follow-up publication the characterization of the ILPR was extended to comprise the tandem repeat sequence of the cytosine-rich strand. NMR spectroscopy and molecular modeling revealed that a synthetic oligonucleotide of the sequence d(CCCCTGTCCCC) could dimerize via formation of intercalating hemiprotonated cytosine–cytosine (C + C) base pairs, the so called *i* motif.[13] It was found that the two d(CCCCTGTCCCC) oligonucleotides associated in an antiparallel fashion with a core of eight C + C base pairs and two symmetric TGT loops, one protruding on each side of the structure (Fig. 2). The arrangement of intercalated C + C base pairs also produces two narrow

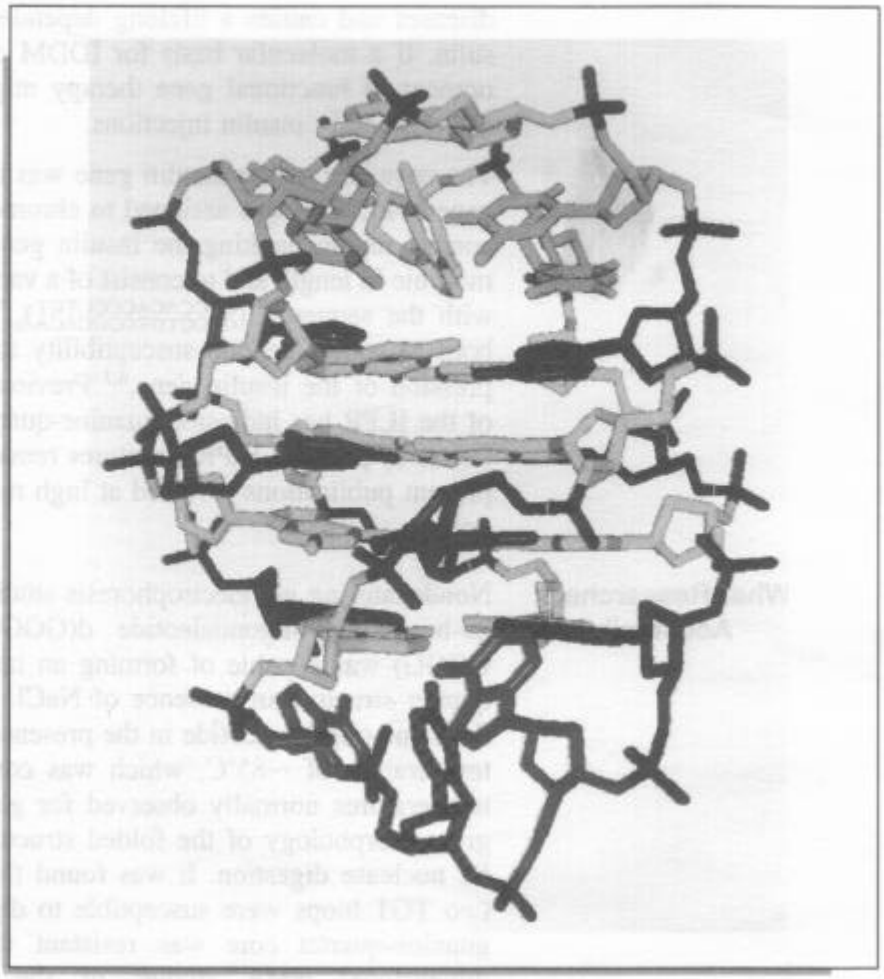

Figure 1. Model showing the intramolecular structure based on four stacked guanine-quartets formed by the guanine-rich strand of the ILPR. The model, which is based on two-dimensional NMR structure determination of d(GGGGTGTGGGGACAGGGGTGTGGGG), has the ACA loop below and the two antiparallel TGT-loops on top of the guanine quartet core. (Reproduced, with permission, from Catasti, P., et al. *J. Mol. Biol.* **1996,** *264,* 534. Copyright © 1996 by Academic Press.)

and two wider grooves; here the cytosines that make up the narrow grooves belong to the same oligonucleotide and are connected by a TGT loop, whereas the cytosines that make up the wider grooves belong to separate oligonucleotides. The stability of the structure originates from electrostatic interactions and stacking of the exocyclic groups rather than the cytidine rings. All nucleotides were found to adopt *anti* glycosidic torsion angles and the majority of the cytosine residues were C3′-endo sugar puckers, whereas the TGT loop nucleotides were C2′-endo puckers.

The 25-base-long oligonucleotide d(CCCCTGTCCCCACACCCCTGTCCCC) is the shortest sequence of the cytosine-rich strand of the ILPR that is capable of forming an intramolecularly folded structure based on the *i* motif. This oligonucleotide was subjected to a pH titra-

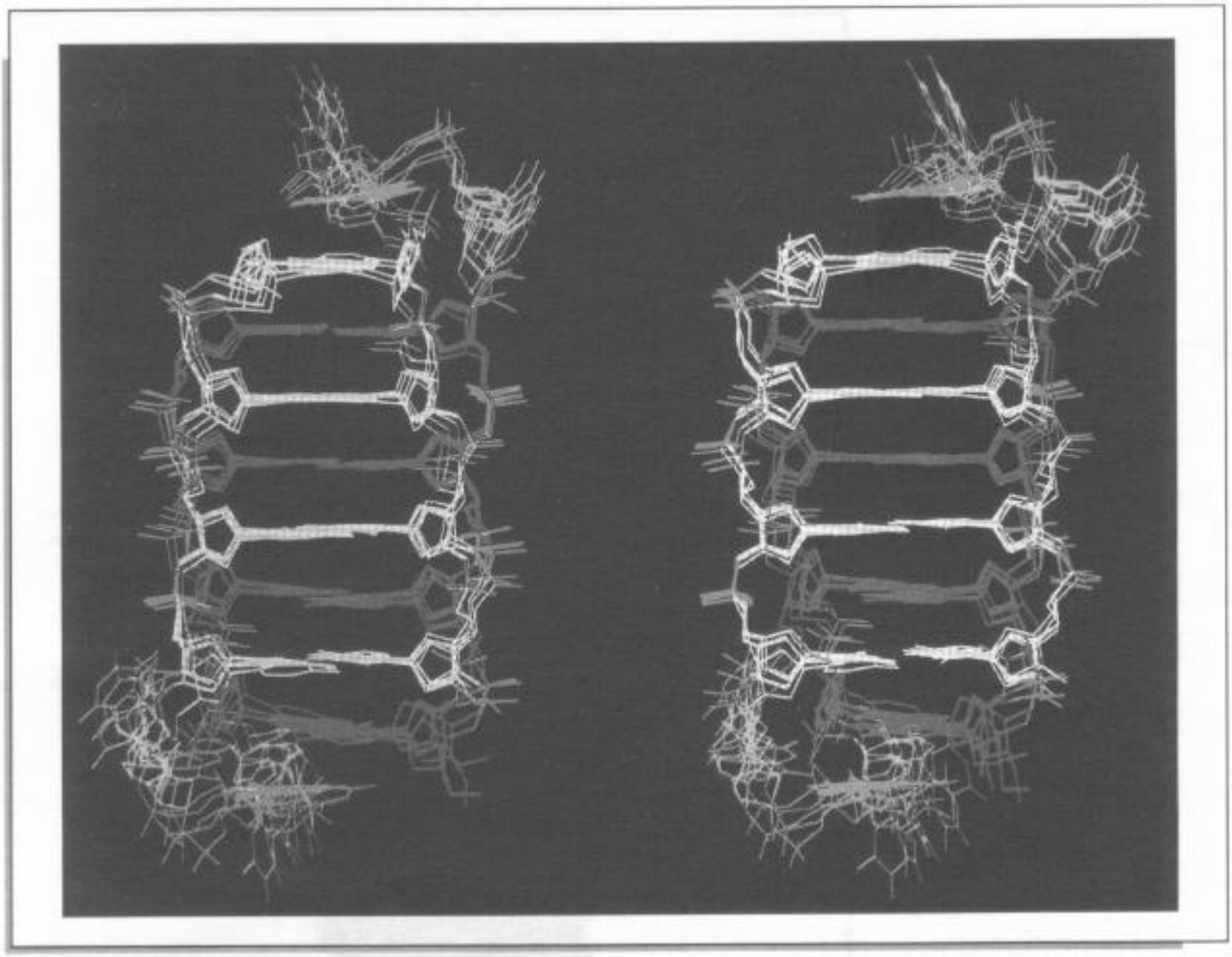

Figure 2. Stereo views of five super-imposed representative structures of the d(CCCCTGTCCCC) dimer shows how an *i* motif consisting of eight intercalated C + C base pairs is formed. One symmetric TGT loop protrudes on each side of the *i* motif. (Reproduced, with permission, from Catasti, P., et al. *J. Mol. Biol.* **1997,** 272, 369. Copyright © 1997 by Academic Press.)

tion monitored by one-dimensional NMR. Imino proton profiles of intercalated C + C base pairs were observed beginning at pH 6.8, and at pH 5.5 the oligonucleotide appeared completely folded. However, the cytosine in the ACA loop exhibited a dual nature: It was present in both protonated and unprotonated states, which indicated the existence of multiple oligonucleotide conformers and prevented two-dimensional NMR structure determination. Notwithstanding, it was concluded that one *i* motif conformer had to be intramolecularly folded with a core of eight intercalated C + C base pairs, two TGT loops on one side, and an ACA loop on the other (Fig. 3).

In vitro DNA replication is arrested whenever a polymerase encounters stable DNA structures in the template strand. This phenomenon has previously been used to indirectly demonstrate guanine-quartet formation in promoter regions.[14] Here, primer extension analyses with single-stranded templates that contained different numbers of the

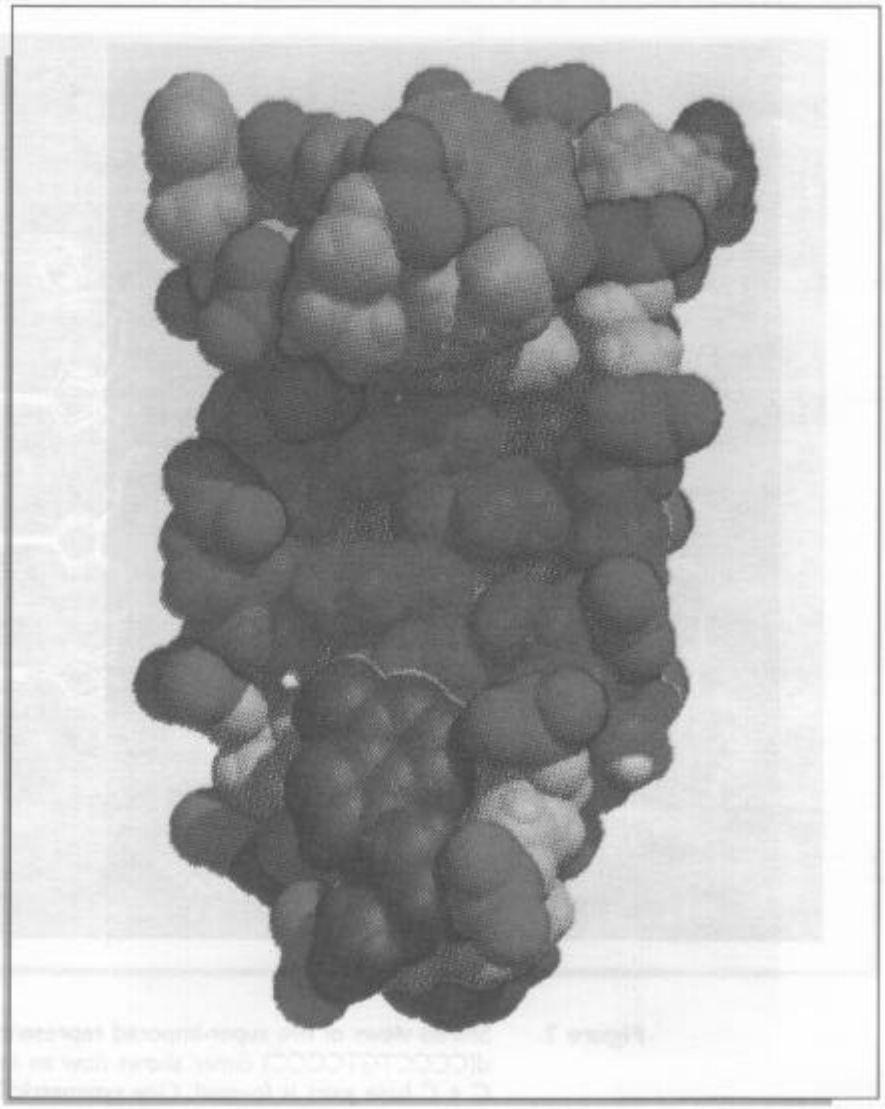

Figure 3. Space-filling model showing the intramolecular structure based on eight intercalated C + C base pairs formed by the cytosine-rich strand of the ILPR. The conformation is based on one-dimensional NMR data for the d(CCCCTGTCCCCACACCCCTGTCCCC) oligonucleotide. One ACA loop is below and the two antiparallel TGT loops are on top of the central *i* motif. (Reproduced, with permission, from Catasti, P., et al. *J. Mol. Biol.* **1997,** *272,* 369. Copyright © 1997 by Acadeic Press.)

cytosine- or guanine-rich repeats were performed. It was discovered that the propensity to form intrastrand structures based on either the guanine-quartet or *i* motif increased with increasing number of $\binom{\text{CCCCACACCCCTGT}}{\text{GGGGTGTGGGGACA}}$ repeats.

Commentary on the Research

Previous Work

The length heterogeneity of the ILPR has been studied in many different individuals and this work has led to the classification into three size-groups according to the number of $\binom{\text{CCCCACACCCCTGT}}{\text{GGGGTGTGGGGACA}}$ repeats: class

I with 26–63 repeats, class II with 64–139 repeats, and class III with 140–200 repeats.[3] It has been established that the insulin gene is expressed bi-allelically in the pancreas and that individuals homozygous for class I predispose strongly for IDDM.[5] Furthermore, in vitro studies have demonstrated that the transcriptional activity of the insulin gene decreases with fewer ($\frac{CCCCACACCCCTGT}{GGGGTGTGGGGACA}$) repeats and that it is independent of the orientation of the ILPR.[7]

Implications In combination with previous research, the two publications condensed here provide a convincing biological role of higher order DNA structure formation in the ILPR. Even though guanine-quartet formation in the ILPR was expected, the model presented gives details concerning loop connectivities, glycosidic torsion angles, and sugar puckering conformations that were previously unavailable. Interestingly, mildly acidic pH was sufficient to induce the *i* motif structure in the complementary cytosine-rich strand, which extends the scope of the *i* motif as a biologically relevant structure. Additionally, the realization that higher order DNA structures form within the ILPR is of utmost importance when searching for transcription factors that regulate insulin expression via ILPR recognition. These will probably be of the single-strand specific type that has been described for other transcription activation mechanisms involving guanine-quartet formation.[15,16]

Summary

A combination of biochemical studies and high-resolution NMR reveals the nature of higher order DNA structures that form in the insulin-linked polymorphic region. The two articles give deeper insight into DNA structure–function relationships in general and provide a basis for better understanding how insulin expression may be regulated at the molecular level.

Further Studies

Although these publications provide a solid foundation to build on, further experimentation is necessary before ILPR-controlled insulin regulation can be fully understood. Here all studies were performed in presence of sodium ions, which preferentially stabilize parallel-stranded guanine-quartet structures.[17] Antiparallel-stranded guanine-quartet structures are stabilized by potassium ions.[12] In live cell nuclei the amount of potassium ions is approximately 30-fold higher than that of sodium ions; normal concentrations are 150 mM potassium and 5 mM sodium. The influence of potassium ions versus sodium ions on guanine-quartet formation in the ILPR needs to be clarified. A number of independent observations lead to the intriguing possibility that switching between different guanine-quartet conformers plays a major role in IDDM: The sodium-potassium ratio affects the equilibrium between parallel and antiparallel guanine-quartet conformations.[18] The complex between the sulfonylurea receptor and inwardly rectifying potassium channels senses the blood glucose level in the pancreas;[19] when glucose is abundant, potassium is actively pumped into the cell and

insulin is secreted. Just like the insulin gene, these two genes are cytogenetically localized to chromosome 11p15[20] and bear guanine-quartet forming motifs in their upstream regulatory regions.[21] To conclude so far, it is plausible that a sodium–potassium switch regulates insulin expression by influencing guanine-quartet forming elements in the regulatory regions of a gene cluster on chromosome 11p15. Finally, guanine-quartet formation may be linked to the autoimmunity of IDDM via the major histocompatibility complex (MHC) genes on chromosome 6p21, which confer additional genetic susceptibility to the disease.[22] The MHC genes control the major pathological feature of IDDM through the T lymphocyte mediated destruction of β cells. A hallmark of MHC genes is their extraordinarily high level of polymorphism. In analogy with the ILPR, one such MHC polymorphic region exhibits variable numbers of consecutive TGGA repeats,[23,24] which constitute perfect motifs for guanine-quartet formation. Future research will show whether guanine-quartet formation in the ILPR, and possibly also in the MHC genes, can explain the genetic susceptibility to IDDM and account for the pathology thereof.

References

1. Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E., Goodman, H.M. *Nature* **1980,** *284,* 26.
2. Owerbach, D., Bell, G.I., Rutter, W.J., Shows, T.B. *Nature* **1980,** *286,* 82.
3. Bell, G.I., Selby, M.J., Rutter, W.J. *Nature* **1982,** *295,* 31.
4. Lucassen, A.M., Julier, C., Beressi, J.P., Boitard, C., Froguel, P., Lathrop, M., Bell, J.I. *Nat. Genet* **1993,** *4,* 305.
5. Bennett, S.T., Lucassen, A.M., Gough, S.C., Powell, E.E., Undlien, D.E., Pritchard, L.E., Merriman, M.E., Kawaguchi, Y., Dronsfield, M.J., Pociot, F., et al. *Nat. Genet.* **1995,** *9,* 284.
6. Lucassen, A.M., Screaton, G.R., Julier, C., Elliott, T.J., Lathrop, M., Bell, J.I. *Hum. Mol. Genet.* **1995,** *4,* 501.
7. Kennedy, G.C., German, M.S., Rutter, W.J. *Nat. Genet.* **1995,** *9,* 293.
8. Hammond-Kosack, M.C., Kilpatrick, M.W., Docherty, K. *J. Mol. Endocrinol.* **1993,** *10,* 121.
9. Hammond-Kosack, M.C., Kilpatrick, M.W., Docherty, K. *J. Mol. Endocrinol.* **1992,** *9,* 221.
10. Hammond-Kosack, M.C., Docherty, K. *FEBS Lett.* **1992,** *301,* 79.
11. Hammond-Kosack, M.C., Dobrinski, B., Lurz, R., Docherty, K., Kilpatrick, M.W. *Nucleic Acids Res.* **1992,** *20,* 231.
12. Sen, D., Gilbert, W. *Methods Enzymol.* **1992,** *211,* 191.
13. Gehring, K., Leroy, J.L., Gueron, M. *Nature* **1993,** *363,* 561.
14. Woodford, K.J., Howell, R.M., Usdin, K. *J. Biol. Chem.* **1994,** *269,* 27029.
15. Simonsson, T., Pecinka, P., Kubista, M. *Nucleic Acids Res.* **1998,** *26,* 1167.
16. Tomonaga, T., Levens, D. *Proc. Natl. Acad. Sci. USA* **1996,** *93,* 5830.
17. Murchie, A.H.I., Aboul-Ela, F., Laughlan, G., Luisi, B., Lilley, D. M. J. *Structure of Parallel-Stranded Guanine Tetraplexes,* Murchie, A.H.I., Aboul-Ela, F., Laughlan, G., Luisi, B., Lilley, D.M.J., eds. Springer: Berlin, 1995, vol. 9, p. 143.
18. Sen, D., Gilbert, W. *Nature* **1990,** *344,* 410.
19. Ammala, C., Moorhouse, A., Gribble, F., Ashfield, R., Proks, P., Smith, P.A., Sakura, H., Coles, B., Ashcroft, S.J., Ashcroft, F.M. *Nature* **1996,** *379,* 545.
20. Inagaki, N., Gonoi, T., Clement, J.P.t., Namba, N., Inazawa, J., Gonzalez, G., Aguilar-Bryan, L., Seino, S., Bryan, J. *Science* **1995,** *270,* 1166.

21. Aguilar-Bryan, L., Nichols, C.G., Wechsler, S.W., Clement, J.P., Boyd, A.E., Gonzalez, G., Herrera-Sosa, H., Nguy, K., Bryan, J., Nelson, D.A. *Science* **1995,** *268,* 423.

22. Davies, J.L., Kawaguchi, Y., Bennett, S.T., Copeman, J.B., Cordell, H.J., Pritchard, L.E., Reed, P.W., Gough, S.C., Jenkins, S.C., Palmer, S.M., et al. *Nature* **1994,** *371,* 130.

23. Saha, B.K., Shields, J.J., Miller, R.D., Hansen, T.H., Shreffler, D.C. *Proc. Natl. Acad. Sci. USA* **1993,** *90,* 5312.

24. Cullen, M., Noble, J., Erlich, H., Thorpe, K., Beck, S., Klitz, W., Trowsdale, J., Carrington, M. *Am. J. Hum. Genet.* **1997,** *60,* 397.