The Quantum Spin Hall Effect

Joseph Maciejko,[1] Taylor L. Hughes,[2] and Shou-Cheng Zhang[1]

[1]Department of Physics, Stanford University, Stanford, California 94305-4060; email: sczhang@stanford.edu

[2]Department of Physics, University of Illinois, Urbana, Illinois 61801

Annu. Rev. Condens. Matter Phys. 2011. 2:31–53

First published online as a Review in Advance on December 3, 2010

The *Annual Review of Condensed Matter Physics* is online at conmatphys.annualreviews.org

This article's doi: 10.1146/annurev-conmatphys-062910-140538

1947-5454/11/0310-0031$20.00

Keywords

topological insulators, spin-orbit coupling, band inversion, mercury telluride, helical edge states

Abstract

Most quantum states of condensed matter are classified by the symmetries they break. For example, crystalline solids break translational symmetry, and ferromagnets break rotational symmetry. By contrast, topological states of matter evade traditional symmetry-breaking classification schemes, and they signal the existence of a fundamentally different organizational principle of quantum matter. The integer and fractional quantum Hall effects were the first topological states to be discovered in the 1980s, but they exist only in the presence of large magnetic fields. The search for topological states of matter that do not require magnetic fields for their observation led to the theoretical prediction in 2006 and experimental observation in 2007 of the so-called quantum spin Hall effect in HgTe quantum wells, a new topological state of quantum matter. In this article, we review the theoretical foundations and experimental discovery of the quantum spin Hall effect.

1. INTRODUCTION

Symmetry appears to be a profound organizational principle of nature. Beautifully symmetric natural patterns are ubiquitous, ranging from spiral galaxies and nearly spherical planets to the fivefold symmetry of starfish and the helical symmetry of DNA. Symmetry also turns out to be a remarkably useful theoretical principle by which we can structure our understanding of complex physical systems. In the field of condensed matter physics—where one studies matter arranged in an endless variety of forms—the symmetry principle may find its most spectacular application. Paradoxically, one of the most instructive ways to study stable phases of matter is to classify them according to the symmetries they break. Most classical and quantum gases and liquids do not break any symmetries; they enjoy the full translational and rotational symmetries of free space. Most classical and quantum solids break these symmetries down to a discrete subgroup of translations and rotations (the space group). Magnetically ordered quantum spin systems such as ferromagnets and antiferromagnets break spin rotation symmetry. This "broken symmetry principle" is at the heart of the phenomenological Ginzburg-Landau theory (1) of phase transitions, which, combined with microscopic many-body theories of condensed matter systems, constitutes the cornerstone of pre-1980s condensed matter physics. In Ginzburg-Landau theory, a stable phase of matter is characterized by a local order parameter, which is nonzero in an ordered phase but vanishes in a disordered phase. Phases with nonzero order parameters are further distinguished by the way the order parameter transforms under symmetry operations, i.e., by the representation of the symmetry group of the system Hamiltonian to which it belongs.

The 1980s were marked by the discovery of the integer (2) and fractional (3) quantum Hall (QH) effects. The QH effect occurs when a two-dimensional electron gas (2DEG), for instance, one formed by electrons trapped in the inversion layer of a metal-oxide-semiconductor structure or electrons in a semiconductor quantum well (QW), is subjected to a large magnetic field perpendicular to the plane of the 2DEG. A perpendicular magnetic field causes the electrons to travel along circular cyclotron orbits, the radii of which becomes smaller with increasing magnetic field. For large enough magnetic fields, electrons in the bulk of the material form small, closed cyclotron orbits. By contrast, electrons near the edge of the sample can trace extended, open orbits that skip along the edge. At low temperatures, quantum effects become important and two series of events happen. First, the area of closed orbits in the bulk becomes quantized, bulk electrons become localized (because they trace only small, closed orbits) and the bulk turns into an insulator. Second, the skipping edge orbits form extended one-dimensional channels with a quantized conductance of e^2/h per channel. Furthermore, the transverse (Hall) conductance σ_{xy} is quantized in integer (integer QH) or rational (fractional QH) multiples of e^2/h.

It was soon realized that the bulk of a QH state is a featureless insulating state that does not break any symmetries other than time-reversal (TR) symmetry and thus cannot be characterized by a local order parameter. Nevertheless, QH states with different values of the Hall conductance are truly distinct phases of matter and correspond to quantum ground states that cannot be adiabatically connected to each other without closing a spectral gap, i.e., without going through a QH plateau transition. Even more surprising is the fact that the quantization of the Hall conductance is extremely accurate even in disordered samples, where one would expect the randomizing effect of disorder to destroy any quantization phenomenon. Indeed, if conduction proceeds only through one-dimensional channels, one would naively expect these to be strongly affected by disorder due to Anderson localization (4).

The lack of a bulk local order parameter description à la Ginzburg-Landau and the existence of boundary states robust to disorder both can be understood as defining characteristics of a topological state of quantum matter. A useful concept in this context is that of bulk-edge correspondence (5), of which the integer QH state provides a clear illustration. A topological state of matter is insulating in the bulk but supports gapless boundary states that are perturbatively robust to disorder. Rather than being characterized by a local order parameter, the bulk is characterized by a topological invariant that, in the case of the integer QH state, is an integer denoted as the TKNN number (6) or Chern number (7). The bulk topological invariant is in turn related to the number of stable gapless boundary states. In the integer QH state, the Chern number is equal to the number of stable gapless edge states and is also the value of the quantized Hall conductance in units of e^2/h. In that sense, one says that the edge states are protected by the bulk topology. But more concretely, what is the mechanism for this "topological protection"? The answer is the following: The bulk topology is responsible for some kind of fractionalization on the edge. More precisely, the usual degrees of freedom of the electron are *spatially separated* on opposite edges. The usual degrees of freedom of an electron in a one-dimensional channel are twofold: right-moving and left-moving. However, in a QH sample, one edge has only right-moving electrons and the other edge has only left-moving electrons (or vice versa, depending on the sign of the magnetic field). Backscattering on a given edge is thus suppressed owing to the inability of an electron to reverse its direction of motion, and the QH edge channels completely evade Anderson localization. Because a single direction of propagation is present on a given edge, the QH edge channels are termed chiral.

The TKNN integer relates the physical response of the Hall conductance to a topological invariant in momentum space. Although the TKNN formalism (6) gives the first insight into the topological nature of the QH state, it is limited to noninteracting systems. A more fundamental description of the QH effect is given by the topological field theory based on the Chern-Simons term in 2+1 dimensions (8, 9). In this approach, the problem of electrons in a 2DEG subject to an external perpendicular magnetic field $\mathbf{B}$ is exactly mapped to that of bosons coupled to both the external magnetic field and an internal, emergent statistical magnetic field $\mathbf{b}$. This statistical magnetic field, the dynamics of which are described by the Chern-Simons term (10), is responsible for the transmutation of the fermionic electrons into bosons. At the magic filling fractions $\nu = 1/m$ (when m is an odd integer) at which the QH effect occurs, the external and statistical magnetic fields precisely cancel each other, and the bosons condense into a superfluid state. The effective field theory of a boson superfluid is the 2+1 Maxwell electrodynamics (11, 12). In the long wavelength and low-energy limit, the Chern-Simons term dominates over the Maxwell term, and the effective theory of the QH state is just the topological Chern-Simons term. This topological field theory is generally valid in the presence of disorder and interactions.

Until very recently, QH states were the only topological states for which the existence had been firmly established by experimental observation. Compared with the rich variety of "traditional" broken-symmetry states, the following question naturally arises: Should there not be other topological states remaining to be discovered? In this article, we review the 2006 theoretical prediction (13) and the 2007 experimental discovery (14) of the quantum spin Hall (QSH) effect in HgTe QWs, a new topological state of matter sharing some similarities—but also several qualitative differences—with the QH effect. We start in Section 2 by reviewing developments in the field of spintronics that took place in the first half of the past decade, as well as theoretical work on the QH effect in the late 1980s, which became key precursor elements in the 2006 theoretical prediction of the QSH effect. We then describe the phenomenology of the QSH state per se. In Section 3, we discuss in greater detail the theoretical prediction of the QSH state in HgTe QWs. In Section 4, we describe the 2007 experimental discovery of the QSH state

in HgTe QWs. In Section 5, we describe the theory of the QSH edge states in greater detail. In particular, we discuss the perturbative stability of the edge states and describe the theoretical prediction of novel phenomena associated with these states, which stem from their unusual electromagnetic response. In Section 6, we give a brief discussion of the recently discovered three-dimensional topological insulators, a three-dimensional generalization of the QSH effect. We then conclude with an outlook on future directions as well as open questions in the field.

2. PHENOMENOLOGY OF THE QUANTUM SPIN HALL EFFECT

One key element that was instrumental in arriving at the theoretical prediction of the QSH state is the prediction of the intrinsic spin Hall (SH) effect in doped semiconductors (15, 16). The SH effect can be thought of as the spin counterpart to the classical "charge" Hall effect. In the SH effect, a transverse spin current flows, say, in the x direction, in response to an applied electric field in the y direction. In contrast to the Hall effect, which breaks TR symmetry due to the applied magnetic field, the SH effect does not break TR symmetry. This can be simply seen by looking at the corresponding response equations. In the Hall effect, the Hall current is given by $J_x = \sigma_{xy}E_y$, where σ_{xy} is the Hall conductivity and E_y is the electric field. Because J_x is odd under TR symmetry but E_y is not, σ_{xy} must necessarily be odd under TR symmetry. Hence, if $\sigma_{xy} \neq 0$, TR symmetry is broken. By contrast, the spin Hall current is given by, say, $J^s_x = \sigma^s_{xy}E_y$, where σ^s_{xy} is the spin Hall conductivity. In contrast to the charge current J_x, the spin current J^s_x is even under TR symmetry (15, 17) and $\sigma^s_{xy} \neq 0$ is consistent with TR symmetry. The role played by the magnetic field in the charge Hall effect is assumed by the spin-orbit coupling of the bandstructure in the SH effect. To theorize a topological state of matter related in some way to the SH effect, one first needs to make sure that the bulk of the system is insulating. The theoretical prediction that the intrinsic SH effect could also be realized in insulators (18) was an important step in that direction.

Because the charge Hall effect naturally leads to the QH effect, asking if the intrinsic SH effect of metals and insulators can similarly have a quantum version follows. Kane & Mele (19) and Bernevig & Zhang (20) independently proposed two systems to realize the QSH effect. Roughly speaking, the QSH state can be viewed as two copies of the QH state with opposite Hall conductances. The proposal by Kane and Mele is based on the spin-orbit interaction of graphene and is mathematically motivated by the earlier work of Haldane (21) on the so-called quantum anomalous Hall effect (QAH effect). The proposal by Bernevig and Zhang is based on the spin-orbit interaction induced by strain in semiconductors. Neither proposal has yet been realized in actual condensed matter systems, mostly because of the small spin-orbit interaction in the proposed systems. However, they provide an important conceptual framework in which the stability of the QSH state can be investigated.

What is quantized in the QSH effect, and in what sense is the QSH state a topological state of matter? These questions are most clearly answered by looking at whether this state supports stable gapless boundary modes, robust to disorder. Let us proceed by comparison with the QH edge modes discussed in Section 1 (**Figure 1**). As mentioned above, the edge states of the QH state are such that electrons can propagate only in a single direction on a given edge. Compared with a one-dimensional system of spinless electrons (**Figure 1**, *top left*), the top edge of a QH system contains only half the degrees of freedom (**Figure 1**, *bottom left*). The QH system can thus be compared to a "freeway" where electrons traveling in opposite directions have to be "driving in different lanes." This spatial separation resulting in chiral edge channels can be illustrated by the symbolic equation $2 = 1+1$ where each 1 corresponds to a different chirality. This "chiral traffic rule" is particularly effective in suppressing electron scattering:

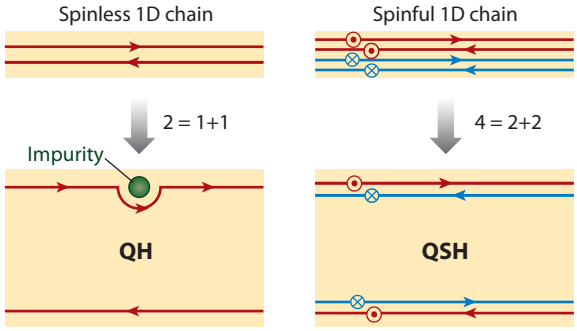

Figure 1

Chiral versus helical: Spatial separation is at the heart of both the quantum Hall (QH) and quantum spin Hall (QSH) effects. A spinless one-dimensional (1D) system (*top left*) has both right-moving and left-moving degrees of freedom. Those two basic degrees of freedom are spatially separated in a QH system (*bottom left*), as illustrated by the symbolic equation 2 = 1+1. The upper edge has only a right-mover and the lower edge a left-mover. These chiral edge states are robust to disorder: They can go around an impurity (*green dot*) without backscattering. By contrast, a spinful 1D system (*top right*) has twice as many degrees of freedom as the spinless system owing to the twofold spin degeneracy. Those four degrees of freedom are separated in a time-reversal (TR) invariant way in a QSH system (*bottom right*). The top edge has a right-mover with spin up (*red dot*) and a left-mover with spin down (*blue cross*), and conversely for the lower edge. That separation is illustrated by the symbolic equation 4 = 2+2. These helical edge states are robust to nonmagnetic disorder, i.e., impurities that preserve the TR symmetry of the QSH state.

Because electrons travel always in the same direction, they are forced to avoid impurities (**Figure 1**, *bottom left*, *green dot*), and thus cannot backscatter.

By contrast, the QSH state can be roughly understood as two copies of the QH state, with one copy for each spin. The edge state structure of the QSH state (**Figure 1**, *bottom right*) can thus be described pictorially by superposing two copies of QH edge states (**Figure 1**, *bottom left*), with opposite chirality for each spin. Compared with a spinful one-dimensional system (**Figure 1**, *top right*), the top edge of a QSH system contains only half the degrees of freedom. The resulting edge states are termed helical, because spin is correlated with the direction of propagation. This new pattern of spatial separation can be illustrated by the symbolic equation 4 = 2+2 where each 2 corresponds to a different helicity. Although electrons are now allowed to travel both forward and backward on the same edge, there is a new "traffic rule" that suppresses backscattering: To backscatter, an electron needs to flip its spin, which requires the breaking of TR symmetry. If TR symmetry is preserved, as is the case for nonmagnetic impurities, no backscattering is allowed (a more detailed discussion of the stability of the QSH edge states and the importance of Kramers's theorem is given in Section 5).

What is the mechanism that allows this spatial separation? In the case of the QH effect, the separation is achieved by an external magnetic field, and in the case of the QAH effect, some internal field breaks TR symmetry. This internal field takes the form of a relativistic mass term for emergent Dirac fermions in 2+1 dimensions, with the sign of the internal field (and hence the chirality of the QAH edge states) dictated by the sign of the mass. In the case of the QSH effect, the separation is achieved through the TR invariant spin-orbit coupling—which is why the QSH insulator can be thought of as an extreme case of the SH insulator discussed previously.

Because the QSH state is characterized by a bulk insulating gap and gapless boundary states robust to disorder (in the presence of TR symmetry), the QSH state is indeed a new topological

state of matter. However, because the Hall conductance of the QSH state vanishes, it is clear that the TKNN or Chern number discussed above, which corresponds to the the value of the Hall conductance in units of e^2/h, cannot provide a useful classification of the QSH state. This issue has been addressed within both the topological band theory (23) and the topological field theory (23). Accordingly, the proper topological invariant is valued in the $\mathbb{Z}_2$ group containing only two elements, 0 or 1, with 1 corresponding to the topologically nontrivial QSH insulator and 0 corresponding to a topologically trivial insulator with no robust gapless edge states. Physically, this $\mathbb{Z}_2$ invariant counts the number of stable gapless edge states modulo 2 (more details in Section 5).

3. THE QUANTUM SPIN HALL EFFECT IN HgTe QUANTUM WELLS

As mentioned above, Kane & Mele (19) proposed graphene—a monolayer of carbon atoms—as a possible candidate for the QSH effect. Unfortunately, this proposal turned out to be unrealistic because the spin-orbit gap in graphene is extremely small (24, 25). The QSH effect was also independently proposed in semiconductors in the presence of strain gradients (20), but this proposal was hard to realize experimentally. Soon afterward, Bernevig, Hughes, and Zhang (BHZ) (13) initiated the search for the QSH state in semiconductors with an "inverted" bandstructure and predicted a quantum phase transition in type-III HgTe/CdTe QW between a trivial insulator phase and a QSH phase governed by the thickness d of the QW. In this section, we review the basic theory of the QSH state in the HgTe/CdTe system. We start by reviewing the basic electronic properties of bulk three-dimensional HgTe and CdTe that make them suitable for hosting the QSH effect (Section 3.1). We then discuss the nature of the two-dimensional subband states in HgTe/CdTe type-III QW and present a simple model that captures the physics of the relevant subbands for the QSH effect (Section 3.2). Finally, we discuss the helical edge states in the HgTe/CdTe QW system in greater detail (Section 3.3).

3.1. Spin-Orbit Coupling and Band Inversion in Bulk HgTe

HgTe and CdTe are binary II-VI semiconductors, both of which crystallize in the zincblende structure. This crystal structure has the same geometry as the diamond lattice, i.e., two interpenetrating face-centered cubic lattices shifted along the body diagonal, but with a different atom on each sublattice. The presence of two different atoms breaks the inversion symmetry of the diamond lattice and reduces the point group symmetry from O_h (cubic) to T_d (tetrahedral). However, the explicit breaking of inversion symmetry has only a small effect on the physics of the QSH state (26). We therefore ignore it from now on and approximate the point group as O_h.

For both HgTe and CdTe, the important bands near the Fermi level are close to the Γ point in reciprocal space and can therefore be indexed according to the Γ-point representations of the cubic group. They are the *s*-type antibonding (parity odd) Γ_6 band and the *p*-type bonding (parity even) band that is split into a $J = 3/2$, Γ_8 band and a $J = 1/2$, Γ_7 band by spin-orbit coupling. CdTe, as shown in **Figure 2*a*** (*right*), is characterized by a band ordering following that of GaAs, with an *s*-type (Γ_6) conduction band and *p*-type (Γ_8,Γ_7) valence bands separated from the conduction band by a large direct energy gap of $\sim$1.6 eV. By contrast, HgTe, as a bulk material, can be regarded as a symmetry-induced semimetal (**Figure 2*a***, *left*). Its negative energy gap of -300 meV indicates that the Γ_8 band, which usually forms the valence band, lies above the Γ_6 band. The light-hole (LH) Γ_8 band becomes the conduction band, the heavy-hole (HH) Γ_8 band becomes the topmost valence band, and the *s*-type Γ_6 band is pushed below the Fermi

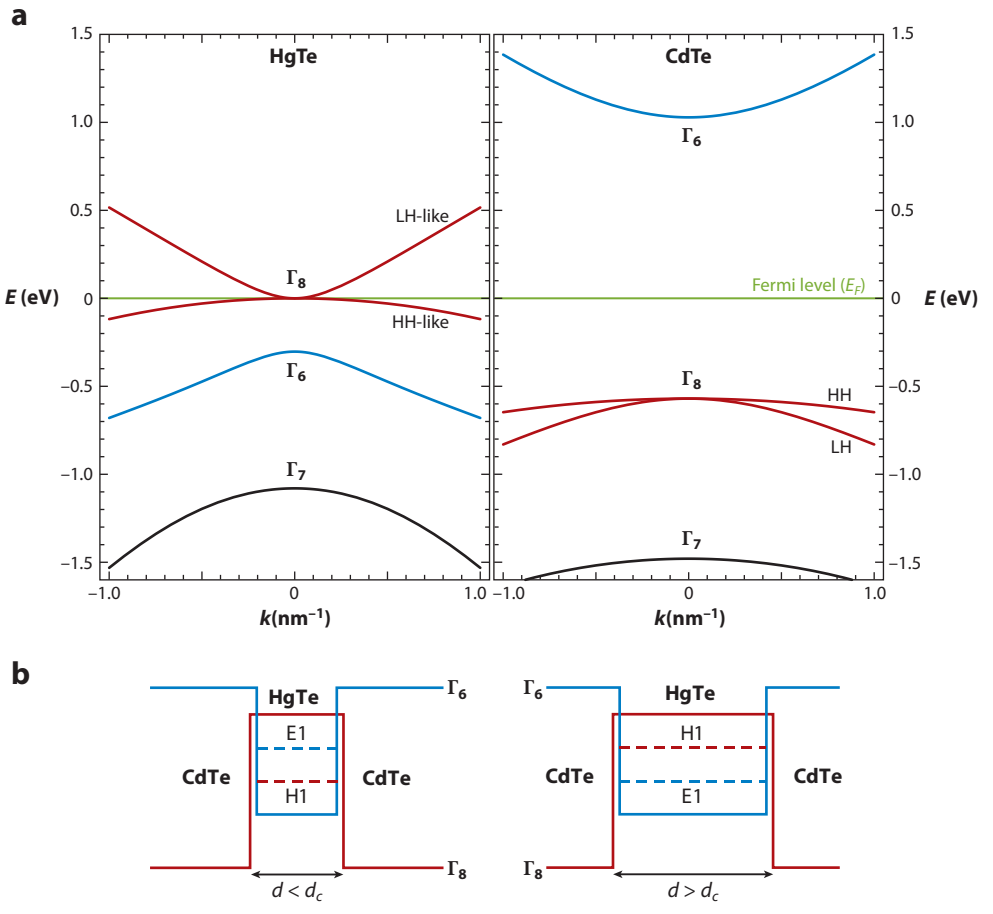

Figure 2

(*a*) Bulk bandstructure for three-dimensional HgTe (*left*) and CdTe (*right*) with Fermi level E_F indicated by a green line; (*b*) schematic picture of the type-III quantum well geometry and lowest-lying subbands for the trivial insulator state with $d < d_c$ (*left*) and the nontrivial QSH insulator state with $d > d_c$ (*right*).

level to lie between the HH band and the spin-orbit split-off Γ_7 band. Because of this unusual sequence of states, such a bandstructure with the associated negative bandgap is termed inverted. Ultimately, because HgTe and CdTe are structurally similar materials with bandgaps of opposite sign, the QSH state can be realized in HgTe/CdTe QWs (discussed in the following sections).

3.2. HgTe Quantum Wells and the Effective Model Hamiltonian

When HgTe-based QW structures are grown, the peculiar properties of the well material can be utilized to engineer the bandstructure in a controlled fashion. More precisely, we discuss the nature of the two-dimensional subbands for the propagation of electrons and holes in the plane perpendicular to the growth axis, which we denote z. The particular QW structure in which we are interested consists of an HgTe QW layer of thickness d sandwiched between thick CdTe barriers (**Figure 2*b***). For wide QW layers (large d), quantum confinement effects are weak

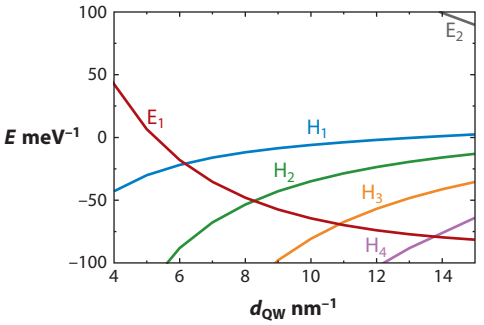

Figure 3

Spectrum of HgTe/CdTe type-III quantum well (QW) subband states as a function of the HgTe QW layer width. The E_1 subband is of mixed Γ_6 and Γ_8 character, and the H_1 subband is of Γ_8 character. For thin wells $d < d_c$, E_1 is the conduction subband, H_1 is the valence subband, and the band ordering qualitatively follows that of bulk "normal" CdTe (**Figure 2***a*, *right*). For thick wells $d > d_c$, the band ordering is reversed and qualitatively follows that of bulk "inverted" HgTe (**Figure 2***a*, *left*).

and the bandstructure remains "inverted," i.e., the physics is mostly "HgTe-like." For the most part, conduction subbands have Γ_8 character and the valence subbands have Γ_6 character. By contrast, for very thin QW layers (small d), the physics is dominated by the neighboring CdTe barriers and the band ordering is "normal," i.e., not inverted. In this case, conduction subbands have mostly Γ_6 character and the valence subbands have mostly Γ_8 character. One therefore expects a critical thickness d_c intermediate to these two regimes, where the Γ_8 and Γ_6 subbands cross and the band ordering changes. This scenario is illustrated in **Figure 3**, where the subband structure has been obtained from self-consistent Hartree calculations using an 8×8 $\mathbf{k}\cdot\mathbf{p}$ model (13, 26, 27). The notation of the subbands as HH (H)-like and electron (E)-like is based on the properties of the respective wave functions (28). The LH-like subbands are far away in energy ($E\lesssim-100$ meV) and are thus not depicted in **Figure 3**. The transition from a normal band alignment to an inverted one can be seen clearly in this figure. For a thin QW layer $d < d_c$, quantum confinement gives rise to a normal subband sequence and the subband gap defined as $E_g \equiv E_{E_1} - E_{H_1}$ is positive. For a thick QW layer $d > d_c$, the subband sequence is inverted and E_g is negative.

Of interest is the derivation of a low-energy effective two-dimensional Hamiltonian to describe the motion of electrons and holes in the plane perpendicular to the growth direction z. In particular, we are interested in QW thicknesses close to the critical thickness d_c. As shown in **Figure 3**, the bands closest in energy to the Fermi energy are the E_1 and H_1 bands. Furthermore, in experiment, the Fermi energy can be tuned by the application of a gate potential to lie in the gap (for $d \neq d_c$) between the E_1 and H_1 bands. The simplest nontrivial effective Hamiltonian will therefore be a Hamiltonian in which only these bands participate. Instead of presenting the details of an explicit calculation, we derive the form this Hamiltonian must take from simple symmetry arguments.

The generic form of the effective Hamiltonian can be inferred from TR symmetry and our assumption of inversion symmetry. We consider a symmetric QW such that even in the QW geometry the inversion symmetry is not broken.[1] This means that the QW wave functions must

[1]Terms that break the structural inversion symmetry (i.e., the reflection $z\to-z$ along the growth direction) but preserve TR symmetry do not destroy the QSH state. In fact, they can be incorporated in the four-band Bernevig-Hughes-Zhang model (see Reference 29).

be eigenstates of the parity operator with eigenvalues ± 1. As such, the E_1 states are odd under parity, whereas the H_1 states are even (13, 27, 28, 29). Furthermore, combining TR symmetry and inversion symmetry, the E_1 and H_1 must both be doubly degenerate. We thus have four basis states $|E_1^+\rangle, |E_1^-\rangle, |H_1^+\rangle, |H_1^-\rangle$ with $\pm$ denoting TR, or Kramers, partners (see Section 5 for a discussion of Kramers's theorem). The states $|E_1^\pm\rangle$ and $|H_1^\pm\rangle$ transform oppositely under parity and a Hamiltonian matrix element that connects them must be odd under parity because we assumed our Hamiltonian preserves inversion symmetry. Thus, to lowest order in the in-plane momentum $\mathbf{k} = (k_x, k_y)$, $(|E_1^+\rangle, |H_1^+\rangle)$ and $(|E_1^-\rangle, |H_1^-\rangle)$ will each be coupled generically via a term linear in $\mathbf{k}$. The $|H_1^+\rangle$ HH state is formed from the spin-orbit-coupled p-orbitals $|p_x + ip_y, \uparrow\rangle$, whereas the $|H_1^-\rangle$ HH state is formed from the spin-orbit-coupled p-orbitals $|-(p_x - ip_y), \downarrow\rangle$, by TR symmetry. Because the total angular momentum J along the z direction is still a good quantum number, these states have $J = 3/2$ and $m_J = \pm 3/2$, respectively. Furthermore, the $|E_1^\pm\rangle$ are formed, roughly speaking, by spin-orbit-coupled s-orbitals $|s, \uparrow\rangle$ and $|s, \downarrow\rangle$ that have $J = 1/2$ and $m_J = \pm 1/2$. Because the $|E_1^\pm\rangle$ and $|H_1^\pm\rangle$ states differ in their total angular momentum J by 1 (in units of $\hbar$), the matrix elements between these states must be proportional to $k_\pm = k_x \pm ik_y$, which carries one unit of angular momentum in the z direction. The only terms allowed in the diagonal elements have even powers of $\mathbf{k}$, including $\mathbf{k}$-independent terms. Because of the twofold degeneracy mentioned above, there can be no matrix elements between the positive (+) state and the negative (−) state of the same band. Finally, the existence of nonzero matrix elements between $|E_1^+\rangle$ and $|H_1^-\rangle$, and similarly for $|E_1^-\rangle$ and $|H_1^+\rangle$, would induce a higher-order process that couples the $\pm$ states of the same band, thus lifting the required degeneracy. Hence, these matrix elements are also forbidden.

If one now writes the effective Hamiltonian according to the symmetry requirements described in the previous paragraph, a striking result is found. This Hamiltonian, first denoted by BHZ, is equal to two copies of the massive Dirac Hamiltonian in 2+1 dimensions (one copy for each spin) with relativistic masses of opposite sign for opposite spins (13). In other words, the HgTe/CdTe QW system precisely realizes the phenomenological QSH model described in Section 2, with two copies of the QAH effect of Haldane for each spin. The BHZ model can thus be considered as the "hydrogen atom" continuum model of the QSH effect. A more detailed calculation (13) reveals that the QSH effect is realized only for $d > d_c$, when the QW bandstructure is inverted, whereas there is no QSH effect (trivial state) for $d < d_c$. The critical thickness is $d_c \simeq 6.3$ nm, in very good agreement with experiments (see Section 4).

3.3. Helical Edge States

As explained in Section 2, we expect that the HgTe/CdTe QW should support gapless helical edge states. To see if the BHZ Hamiltonian does indeed predict such edge states, we study this Hamiltonian on a strip of finite width, say, in the y direction. This problem has been studied both numerically (26, 30) and analytically (26, 30, 31), and edge states are indeed found when the system is in the QSH phase, i.e., for $d > d_c$ (**Figure 4**).

The edge state properties can be determined from the edge state wave functions, which are obtained numerically or analytically (26, 30, 31). First, the edge states are exponentially localized on the edge. The associated decay length ξ is roughly given by $\xi \sim \hbar v/E_g$, where E_g is the E_1-H_1 gap mentioned above and v is the velocity of the edge states (nominally the velocity of the bulk Dirac point). Second, the edge states are indeed helical, as anticipated by the phenomenological model of the QSH state described in Section 2. At energies smaller than E_g, the edge states disperse linearly and can be described by a one-dimensional Dirac equation. However, the simple picture of a chiral edge state for spin up and an antichiral edge state for

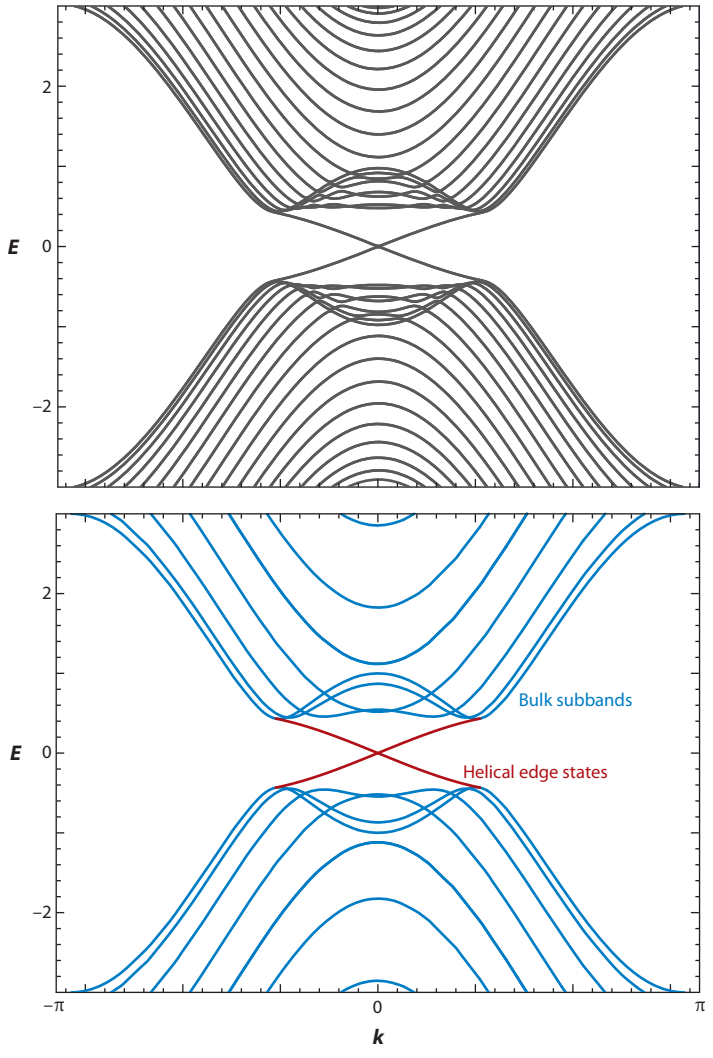

Figure 4

Calculated quasi-one-dimensional subband dispersion for a strip of HgTe/CdTe quantum well of finite width in the *y* direction, with *k* the momentum along the *x* direction: (*top*) exact numerical diagonalization of the tight-binding form of the Bernevig, Hughes, and Zhang Hamiltonian; (*bottom*) analytical solution with bulk subbands in blue and helical edge states in red. One can clearly see the helical edge states dispersing linearly within the E_1-H_1 gap.

spin down is not rigorously exact. As described in Section 3.2, the QW states are strongly spin-orbit coupled and thus are not eigenstates of the spin operator, but rather of the total angular momentum in the *z* direction. Therefore spin is not conserved. Furthermore, in the presence of a boundary, even the conservation of total angular momentum in the *z* direction fails. However, TR symmetry is preserved, and Kramers's theorem still holds. The defining characteristic of the helical edge state is that the two states with opposite chirality on a given edge transform into each other under TR, forming what is known as a "Kramers pair." This property does not require any symmetry other than TR and is robust under the introduction of disorder

(Section 5). This is useful to keep in mind, as many theoretical works in the field use the simple "spin up/spin down" picture but really mean Kramers partners. This is also the reason why the QSH effect is not to be understood as a quantized SH effect: Because spin-orbit coupling destroys spin conservation, there is no such thing as a quantized SH conductance in the QSH effect. This is another way to understand why the correct topological invariant for the QSH effect is $\mathbb{Z}_2$ and not $\mathbb{Z}$. Finally, the BHZ Hamiltonian predicts a single helical edge state per edge. This is useful when we compare the theoretical predictions of the BHZ model to experiment in Section 4.

4. EXPERIMENTS ON HgTe QUANTUM WELLS

Less than one year after the 2006 theoretical prediction described in Section 3, a team at the University of Würzburg led by Laurens W. Molenkamp observed the QSH effect in HgTe/CdTe QWs grown by molecular beam epitaxy (14). In this section, we review the main results of these experiments.

4.1. Landau Levels and Band Inversion in HgTe Quantum Wells

As described in Section 3, the QSH effect relies heavily on the existence of band inversion in bulk HgTe and its consequences for the HgTe/CdTe QW subband structure. Therefore, one should first verify whether band inversion in the HgTe/CdTe system exists. A striking manifestation of this is a so-called re-entrant QH effect (26) that has been experimentally observed (Section 1.3) (see **Figure 5**). The peculiar band structure of inverted HgTe/CdTe QWs gives rise

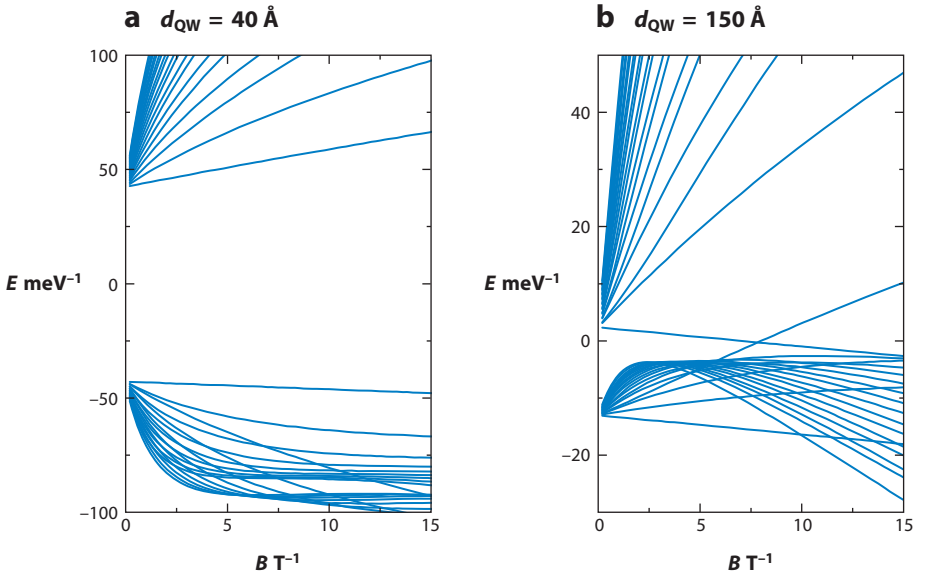

Figure 5

Bulk Landau levels (*fan diagram*) for an HgTe/CdTe quantum well (QW) in a perpendicular magnetic field B. (*a*) Trivial insulator ($d < d_c$): No level crossing occurs as a function of B, and for a fixed Fermi energy E_F in the $B = 0$ gap, the Hall conductance σ_{xy} is always zero. (*b*) Quantum spin Hall insulator with $d > d_c$: There is a level crossing at some critical field $B = B_c$, and for a fixed E_F in the $B = 0$ gap, a conduction or valence band Landau level eventually crosses E_F, giving rise to a re-entrant quantum Hall effect with $\sigma_{xy} = \pm e^2/h$.

to a unique Landau level (LL) dispersion. For a normal band structure ($d < d_c$), all LLs are shifted to higher energies for increasing magnetic field B (**Figure 5*a***). This is the usual behavior and can be commonly observed for a variety of semiconductors. When the band structure of the HgTe/CdTe QW is inverted ($d > d_c$), the LL dispersion is markedly different (**Figure 5*b***). Due to the mixing of electron- and hole-like states, one of the states of the H_1 conduction subband is a pure HH state ($m_J = -3/2$). Consequently, the energy of the corresponding LL decreases with increasing B. In addition, one of the E_1 valence subband LLs has mainly electron character and thus shifts to higher energies with increasing B. This leads to a crossing of these two peculiar LLs at some critical value of the magnetic field $B = B_c$. The existence of such an LL crossing is a clear indication of the occurrence of an inverted band structure.

The crossing of the conduction and valence subband LLs for a QSH insulator ($d > d_c$) can be verified experimentally by QH experiments. We consider that the Fermi level E_F is fixed and lies inside the zero-field gap $E_g(B = 0)$. At $B = 0$, the Hall conductance vanishes $\sigma_{xy} = 0$. At small enough fields, E_F is still in the gap; therefore, a vanishing σ_{xy} remains. For large enough fields, however, one of the "peculiar" LLs will cross E_F. Whether a conduction or a valence subband LL crosses E_F first depends on the position of E_F and on the exact slope of the LL dispersions (**Figure 5*b***). In any event, an LL will cross the Fermi level and will therefore give rise to a re-entrant QH effect with $\sigma_{xy} = -e^2/h$ (for a valence LL) or $\sigma_{xy} = e^2/h$ (for a conduction LL). For even larger B, the second "peculiar" LL will cross E_F and "cancel" the re-entrant QH state, restoring the initial vanishing Hall conductance $\sigma_{xy} = 0$. This re-entrant QH effect has been experimentally observed (14), which confirms the phenomenon of band inversion in HgTe/CdTe QWs.

4.2. Transport Measurements: Edge Conductance and Nonlocal Transport

However, the observation of band inversion alone does not constitute a discovery of the QSH effect. The most striking feature of the QSH state may be the existence of protected gapless edge states (Section 3.3), so one would like to observe these experimentally. Because, as discussed in Section 3.3, we do not expect any quantized SH conductance, the most natural experiment involves the observation of the transport of charge by these edge states. However, because the QSH effect exists in the absence of an external magnetic field, one cannot perform a measurement of the Hall conductance, which would be zero. The simplest measurement is thus to measure the longitudinal conductance G on a strip of HgTe/CdTe QW of finite width. It is well known (32) that a single quantum channel has a longitudinal conductance of e^2/h. Because the BHZ model (Section 3.2) predicts a single helical edge state per edge, and a strip has two edges, we expect a longitudinal conductance $G = 2e^2/h$, independent of the width and length of the sample. Indeed, there is a right-mover (say) on each edge. By comparison, a QH system on a strip with a single chiral edge state per edge would give $G = 2e^2/h$, because only one of the edges has a right-mover. In **Figure 6** we show the experimental data (14) obtained for an HgTe/CdTe QW in the inverted regime $d > d_c$. The longitudinal resistance reaches a plateau $R_{xx} \simeq h/2e^2$ for values of the Fermi level inside the gap (the Fermi level can be adjusted by the gate voltage V_g). This quantization has been observed for several samples and at various temperatures. More precisely, the longitudinal resistance is independent of sample width and length, provided the sample size is smaller than the phase coherence length. As discussed in the following sections, unlike the QH effect, the stability of the QSH edge states relies on Kramers's theorem, which assumes phase coherence.

These first experimental results provide strong evidence for the existence of the QSH state, by confirming the basic predictions of the BHZ model. One can submit the edge state prediction

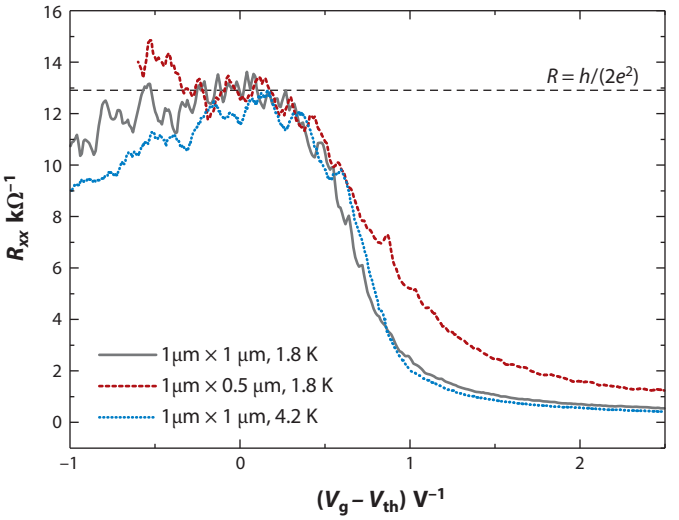

Figure 6

Measurement of the longitudinal resistance R_{xx} on strips of HgTe/CdTe quantum well in the inverted regime ($d > d_c$). The gate voltage V_g is adjusted to position the Fermi level inside the gap ($V_g - V_{th} \simeq 0$). The resistance reaches a quantized plateau $R_{xx} \simeq h/2e^2$ indicated by the horizontal line, for small enough samples and at low enough temperatures where quantum phase coherence is preserved.

of the BHZ model to even more stringent tests by performing multiterminal transport experiments (33; see also 34). In these experiments, one measures various nonlocal resistances denoted $R_{ij,kl}$ by passing a current from terminal i to terminal j and detecting a voltage between terminals k and l. Within the Landauer-Büttiker formalism of phase-coherent transport (32, 35), these resistances can be expressed in terms of a transmission matrix $T_{ij} \equiv T_{i \leftarrow j}$ giving the probability for an electron to transmit from terminal j to terminal i. For a general two-dimensional sample, the number of transmission channels scales with the width of the sample, such that the transmission matrix T_{ij} is complicated and nonuniversal. However, a tremendous simplification arises if the quantum transport is dominated by edge states. Consider labeling consecutive terminals of an N-terminal device such as a Hall bar (**Figure 7**, *insets*) by consecutive integers $i = 1, 2, \ldots, N$. In the $\nu = 1$ QH effect, there is one chiral edge state going along the boundary from each terminal i to (say) its neighbor on the right $i+1$, but not to its neighbor on the left $i-1$ (because the edge state is chiral) nor to any other terminal. Therefore, the transmission matrix for the QH state reads $T^{QH}_{i+1,i} = 1$ for all i and is zero otherwise. From this simple transmission matrix, one can solve the Landauer-Büttiker equations (35, 36), with results such as a vanishing four-terminal resistance $R_{14,23} = 0$ and a finite two-terminal resistance $R_{14,14} = h/e^2$ for the $\nu = 1$ QH effect. In contrast, in the QSH effect, the edge states are helical and consist of counterpropagating Kramers partners. As explained above, this means that we can consider the helical edge states as two copies of chiral edge states related by TR symmetry, and the transmission matrix follows as $T^{QSH} = T^{QH} + (T^{QH})^{\dagger}$. Therefore, we obtain $T^{QSH}_{i+1,i} = T^{QSH}_{i,i+1} = 1$ and all other T^{QSH}_{ij} vanish. One can again solve the Landauer-Büttiker equations for the four- and two-terminal resistances taken as examples above, and we obtain $R_{14,23} = h/2e^2$ and $R_{14,14} = 3h/2e^2$. These resistances are quantized but manifestly different from the QH case. Such nonlocal transport measurements have been performed (**Figure 7**) with very good agreement with the theoretical Landauer-Büttiker predictions in the helical edge state

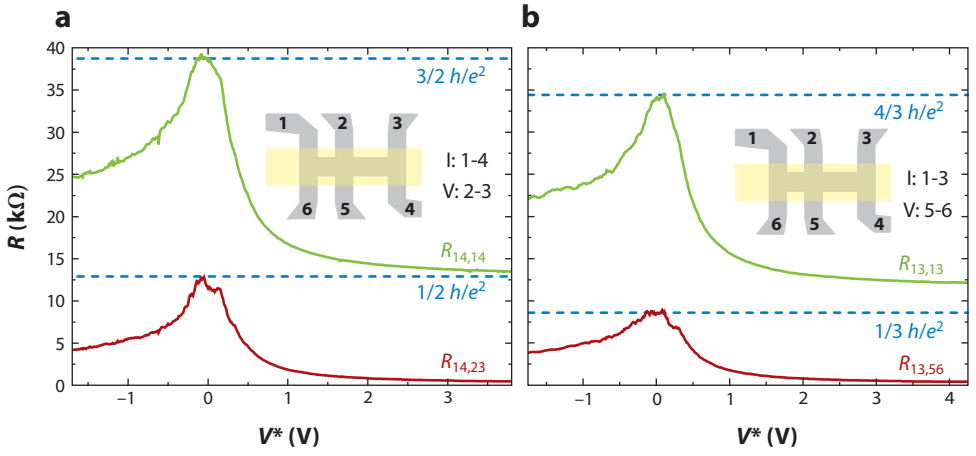

Figure 7

Measurement of the nonlocal resistances $R_{ij,kl}$ in an HgTe/CdTe Hall bar. When the sample is gated into the quantum spin Hall regime ($V^* \simeq 0$), the nonlocal resistances follow closely the predictions of the Landauer-Büttiker formalism for helical edge states.

picture. These and other nonlocal transport measurements in different multiterminal geometries (33) can be taken as definitive evidence for the existence of edge channel transport in the QSH regime.

5. THEORY OF THE HELICAL EDGE STATE

In the above sections, we allude several times to the fact that helical edge states should be robust to disorder owing to the bulk topology of the QSH state and the fact that this bulk topology is protected by a discrete symmetry, TR symmetry. We now explain this assertion in more detail.

The generic properties of TR symmetry are important for understanding the properties of the helical edge states. The antiunitary TR operator has different properties depending on whether the degrees of freedom considered carry integer or half-odd-integer angular momentum. For half-odd-integer angular momentum, we have $T^2 = -1$, which implies, according to a theorem by Kramers, that any eigenstate of a single-particle Hamiltonian must have a degenerate partner. In addition, matrix elements of a single-particle TR invariant perturbation between a state and its Kramers partner must vanish identically (22, 37). As a result, TR symmetry forbids scattering or hybridization between a state and its Kramers partner. For integer angular momentum, we have $T^2 = +1$, and there are no constraints on the single-particle energy spectrum.

The robustness of the QSH edge state is a direct consequence of the property $T^2 = -1$ for half-odd-integer angular momentum. In fact, the counterpropagating states on the same edge are Kramers partners. Therefore, no single-particle TR invariant perturbation can backscatter the helical edge states, and each edge will carry a longitudinal conductance of e^2/h per helical edge state. Furthermore, for a translationally invariant system, Kramers's theorem ensures the crossing of the edge state dispersions at special points in the Brillouin zone (**Figure 4**, edge state crossings at $k = 0$). Because of this level crossing, the spectrum of a QSH insulator cannot be adiabatically continued into that of a topologically trivial insulator without helical edge states.

This is the starting point for the definition of the $\mathbb{Z}_2$ invariant (22, 23, 37–40). The $\mathbb{Z}_2$ nature of the invariant can be easily understood in the following way: If there is more than one helical edge state on each edge, edge states may annihilate in pairs without violating TR symmetry. Thus, if an odd number of helical edge states is present initially, then at least one helical edge state will remain after such annihilation processes and the state will remain topologically nontrivial. If an even number of helical edge states is present initially, then all edge states can annihilate pairwise and a topologically trivial insulating state with no edge states results. This even-odd distinction is precisely the reason why the invariant is $\mathbb{Z}_2$ valued. By contrast, if TR symmetry is broken, say, by a magnetic field or a magnetic impurity, backscattering is not forbidden and the gaplessness of the edge states is not protected. In this case, a gap will generally open in the edge state dispersion. In the language of one-dimensional Dirac fermions on the edge, the gap corresponds to a relativistic mass term, and we use the terms gap and mass interchangeably.

A nice semiclassical picture illustrates why single-particle backscattering is forbidden for degrees of freedom with half-odd-integer angular momentum (**Figure 8**). The mechanism is analogous to the way antireflective coatings on eyeglasses and camera lenses work. In such a system, reflected light from the top and bottom surfaces of the antireflective coating interfere destructively, suppressing the overall amount of reflected light (**Figure 8*a***). This effect is, however, not robust, as it requires precise matching of the coating thickness to the wavelength of the light. Just as photons can be reflected from an interface between two dielectrics, so can electrons be backscattered by an impurity, and different backscattering paths will interfere with each other (**Figure 8*b***). On a QSH edge, the two paths correspond to the electron going around the impurity in either a clockwise or counterclockwise fashion, with the spin rotating by an angle of π or $-\pi$, respectively. Consequently, the phase difference between the two paths is a full 2π rotation of the electron spin. However, the wave function of a spin-1/2 particle picks up a minus sign under a full 2π rotation. Therefore, the two backscattering paths related by TR always interfere destructively, leading to perfect transmission. In contrast to the antireflective

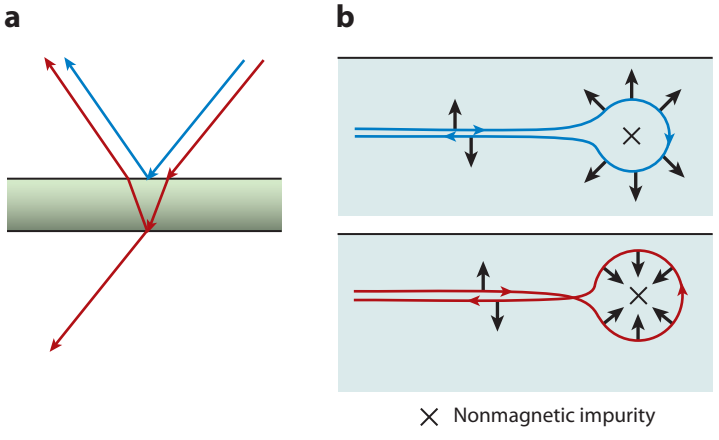

Figure 8

(*a*) On a lens with an antireflective coating, light beams reflected from the top (*blue*) and bottom (*red*) surfaces interfere destructively, which suppresses the overall amount of reflected light. (*b*) Two possible paths taken by a quantum spin Hall (QSH) edge electron when scattered by a nonmagnetic impurity. The spin is rotated by 180 degrees clockwise along the blue curve and counterclockwise along the red circle. A geometrical phase factor associated with the rotation of the spin leads to destructive interference between the two paths, leading to the suppression of electron backscattering on the QSH edge.

coating, this effect is robust (at the single-particle level), as it depends only on the basic property $T^2 = -1$ for half-odd-integer spin particles.

5.1. Stability of the Helical Liquid: Disorder and Interactions

As discussed in Section 4, we expect the QSH edge states to be robust under single-particle perturbations that preserve TR symmetry. This includes isolated nonmagnetic impurities and, more generally, quenched nonmagnetic disorder (37, 38, 41, 42) [in the presence of a external magnetic field that breaks TR symmetry, even nonmagnetic disorder can localize the QSH edge states (43, 44)]. To be more precise, we consider the case of a very wide sample, i.e., $W \gg \xi$, where W is the sample width and ξ the decay length of the edge states. If a perturbation can scatter an electron from one edge of the sample to the other, it can gap the edge states even in the presence of TR. We are, however, not interested in such nonlocal perturbations, but rather, our interest lies in perturbations that are local to the edge. That being said, we have so far discussed only the QSH edge states in the absence of interactions. Are the QSH edge states stable to interactions?

If interactions are allowed, there exist processes that can backscatter electrons on the QSH edge without violating TR symmetry. The simplest of these is a correlated two-particle backscattering process (37, 38, 45, 46), in which two right-moving (for example) spin-up electrons are backscattered as two left-moving spin-down electrons. Why this process preserves TR symmetry can be understood in a simple way using the semiclassical picture of electron backscattering described in the previous section and in **Figure 8*b***. Whereas the wave function of a single electron will pick up a minus sign under a 2π spin rotation, the wave function of two electrons backscattered in the same way will pick up two minus signs that cancel each other. Therefore, the two time-reversed processes interfere constructively, and two-particle backscattering (or more generally, $2n$-particle backscattering with n integer) is allowed by TR symmetry (37, 38, 45, 46).

However, the QSH edge state is perturbatively stable to electron-electron interactions, in the sense that a gap will be opened only if the interaction strength exceeds a critical value. In one-dimensional systems such as the QSH edge, the strength of (short-range) electron-electron interactions can be parametrized by a number K called the Luttinger parameter, where $K = 1$ corresponds to noninteracting electrons, $K > 1$ to attractive interactions, and $0 < K < 1$ to repulsive interactions. For repulsive electron-electron interactions, perturbation theory in the two-particle backscattering amplitude (38) shows that the QSH edge states become gapped if $K < K_c$, where $K_c = 1/2$.. This result holds for a uniform (umklapp) two-particle backscattering perturbation. If the two-particle backscattering happens only at a single point along the edge (38), as is the case for an impurity-induced process, the critical value is reduced to $K_c = 1/4$, i.e., even stronger interactions. Finally, in the case of quenched disorder-induced random two-particle backscattering (38, 42), the critical interaction strength is $K_c = 3/8$. Although these values correspond to rather strong electron-electron interactions, numerical estimates (46, 47) of the Luttinger parameter show that one could possibly achieve $K \sim 0.5$ in the HgTe/CdTe system or even $K \sim 0.2$ in the recently proposed InAs/GaSb/AlSb type-II QW structure (48), which should also realize the QSH effect.

Another perturbation that can potentially gap the QSH edge states is a magnetic impurity. On the one hand, classical magnetic impurity acts as a local Zeeman field, which splits the degeneracy between spin up and spin down and opens a gap. On the other hand, a quantum magnetic impurity, i.e., a Kondo impurity, exhibits subtler behavior. The physics of the Kondo effect is that of a crossover from a high-temperature regime at $T \gg T_K$ where the magnetic

impurity (say, a spin-1/2 impurity) acts as a free spin decoupled from the edge eletrons, to a low-temperature regime at $T \ll T_K$ where the impurity spin is screened by the edge electrons (49). T_K is the Kondo temperature that represents the crossover scale for this problem. For $T \gg T_K$, the QSH edge electrons will be weakly backscattered by spin flips of the impurity. As T approaches T_K, spin flips are more frequent and the edge conductance decreases (46). However, at even lower temperatures $T \ll T_K$, the impurity spin is screened and behaves effectively as a spinless, nonmagnetic impurity (38, 46, 49). Therefore, the conductance must be restored to e^2/h per edge at $T = 0$, owing to the helical nature of the QSH edge (46). In general, the interplay between interactions and the helical nature of the QSH edge states gives rise to qualitatively different transport properties as compared with ordinary one-dimensional quantum wires or QH edge states (47, 50–54).

5.2. Fractional-Charge Effect and Spin-Charge Separation

Although a number of interesting physical phenomena associated with the QSH effect have been mentioned above, such as nonlocal edge transport, suppression of backscattering, and an "anomalous" Kondo effect, we have not yet fully addressed the question of how the $\mathbb{Z}_2$ invariant can be measured, although we have discussed indirect consequences of the bulk topology. We now discuss two physical properties of the QSH state directly related to the $\mathbb{Z}_2$ invariant, which can, in principle, be measured experimentally.

The first property we discuss is the existence of a localized fractional charge at the edge of a QSH sample when a magnetic domain wall is present (55). As explained above, a magnetic field (or ferromagnetic layer) generates a mass term for the one-dimensional edge Dirac fermions; hence, we can speak of a mass domain wall. The idea of fractional charges in condensed matter systems induced on a mass domain wall goes back to the Su-Schrieffer-Heeger model of polyacetylene (56), where the mass there corresponds to a charge density wave order parameter. For spinless fermions, a mass domain wall induces a localized state with charge $e/2$. The continuum field theory description of this problem corresponds to the Jackiw-Rebbi soliton (57). However, for a real material such as polyacetylene, two spin states are present for each electron. Because of this doubling, a mass domain wall in polyacetylene carries only integer charge. Indeed, conventional one-dimensional electron systems such as polyacetylene have four basic degrees of freedom, i.e., right- and left-movers with two spin orientations. However, the helical edge state on a given edge of the QSH insulator has only two degrees of freedom: a spin-up right-mover and a spin-down left-mover. Therefore, the QSH helical edge state has half the degrees of freedom of a conventional one-dimensional system and thus avoids the doubling problem. Because of this fundamental topological property of the helical liquid, a magnetic domain wall carries $e/2$ charge (55). In addition, if the magnetization is rotated adiabatically (i.e., with angular frequency $\omega \ll E_g/\hbar$), a quantized current will flow, with a quantized charge e pumped after each cycle. This provides a direct realization of the Thouless pump (58). The fractional charge effect is also realized in the presence of two-particle backscattering at an impurity site for strong electron-electron interactions $K < 1/4$, where instanton effects at low temperatures correspond to tunneling of excitations with $e/2$ charge (46).

Although the fractional charge effect is truly a topological effect, it still occurs only on the edge. Because the QSH effect is defined in terms of bulk topology, it would be satisfying to have bulk physical observables that directly probe this topology. Such an observable is given by the spin-charge separation effect (59, 60). We first adopt the simple picture of the QSH state with S^z conservation as two copies of the QH state for opposite spins and then comment on the realistic case of no spin conservation. The idea essentially follows the Laughlin gauge argument

(61) for the QH effect. We consider threading adiabatically a thin $hc/2e$ magnetic flux tube (π flux) through the bulk of a QSH sample. Because both spin species carry the same charge, electrons of both spins feel the same flux $\varphi_\uparrow = \varphi_\downarrow = \pi$. We now consider a Gauss loop surrounding the flux tube. As the flux $\varphi_\uparrow$ is turned on adiabatically from 0 to π, Faraday's law of induction states that a tangential electric field $\mathbf{E}_\uparrow$ is induced along the Gauss loop. The quantized Hall conductance for spin-up electrons implies a radial current $\mathbf{j}_\uparrow = \frac{e^2}{h}\hat{\mathbf{z}} \times \mathbf{E}_\uparrow$, resulting in a net charge flow $\Delta Q_\uparrow = e/2$ (62, 63) and a net spin flow $\Delta S^z_\uparrow = \hbar/4$ (64) through the Gauss loop when $\mathbf{j}_\uparrow$ is integrated over time. An identical argument applied to the spin-down component yields $\mathbf{j}_\downarrow = -\frac{e^2}{h}\hat{\mathbf{z}} \times \mathbf{E}_\downarrow$, $\Delta Q_\downarrow = -e/2$, and $\Delta S^z_\downarrow = \hbar/4$. Therefore, this process creates a state with total charge $\Delta Q = 0$ and total spin $\Delta S^z = \hbar/2$, i.e., a spinon (59, 60). Because a flux of π is equivalent to a flux of $-\pi$ owing to the two compact $U(1)$ symmetries (charge and spin S^z), we can also formally insert a spin flux $\varphi_\uparrow = -\varphi_\downarrow = \pi$, which gives rise to a holon state with $\Delta Q = -e$ and $\Delta S^z = 0$, or a spin flux $\varphi_\uparrow = -\varphi_\downarrow = -\pi$, which gives rise to a chargeon state with $\Delta Q = e$ and $\Delta S^z = 0$. A holon (chargeon) is a spinless particle with negative (positive) electric charge. In the absence of S^z conservation, one can still define generalized spinon and holon/chargeon states solely in terms of their transformation properties under TR (59, 60). Although not dynamical excitations, these spin-charge-separated soliton states provide a striking physical consequence of the bulk topology of the QSH insulator, and they can be used to provide a bulk definition of the $\mathbb{Z}_2$ invariant beyond topological band theory (59, 60).

6. TOPOLOGICAL INSULATORS IN THREE DIMENSIONS

From the above discussions, we see that the simplest TR invariant two-dimensional topological insulator, the QSH insulator in HgTe/CdTe QWs, has an insulating gap in the bulk and one pair of helical edge states at each edge. A topological phase transition occurs as a result of the band inversion at the Γ point driven by the spin-orbit interaction. The helical edge state forms a single one-dimensional massless Dirac fermion with counter-propagating states forming a Kramers doublet under TR symmetry. Furthermore, the helical state consisting of a single massless Dirac fermion is "holographic," in the sense that it cannot exist in a purely one-dimensional system, but can exist only as the boundary of a two-dimensional system (38).

The model Hamiltonian for the two-dimensional topological insulator in HgTe/CdTe QWs also gives a basic template for a generalization to three dimensions, leading to a simple model Hamiltonian for a class of materials: Bi_2Se_3, Bi_2Te_3, and Sb_2Te_3 (65, 66). Similar to their two-dimensional counterpart the HgTe/CdTe QW, these materials can be described by a simple but realistic model where the spin-orbit interaction drives a band inversion transition at the Γ point. In the topologically nontrivial phase, the bulk states are fully gapped, but there is a topologically protected surface state consisting of a single two-dimensional massless Dirac fermion. This two-dimensional massless Dirac fermion is "helical," as the spin of the electron points perpendicularly to its momentum, forming a left-handed helical texture in momentum space. Similar to the one-dimensional helical edge state, a single two-dimensional massless Dirac surface state is "holographic," in the sense that it cannot occur in a purely two-dimensional system with TR symmetry but can exist as the boundary of a three-dimensional insulator. A TR invariant single-particle perturbation cannot introduce a gap for the surface state. A gap can open up for the surface state when a TR breaking perturbation is introduced on the surface. In this case, the system becomes a full insulator, both in the bulk and on the surface. The topological properties of the fully gapped insulator are characterized by a novel topological magnetoelectric effect (23).

Soon after the theoretical prediction of the three-dimensional topological insulator in the Bi_2Se_3, Bi_2Te_3, Sb_2Te_3 class of materials (65, 67), angle-resolved photoemission (ARPES) experiments demonstrated the surface state with a single Dirac cone (67–69). Furthermore, spin-resolved ARPES experiments observed the left-handed helical spin texture of the massless Dirac fermion (69). These pioneering theoretical and experimental works inspired much of the subsequent developments both in theory and experiment.

The general theory of the three-dimensional topological insulator has been developed along two different routes. The topological band theory gives a general description of the topological invariant in the single-particle momentum space (39, 70, 71). In particular, a method due to Fu & Kane (72) gives a simple algorithm to determine the topological properties of any complex electronic structure with inversion symmetry. This method predicts that the semiconducting alloy Bi_xSb_{1-x} is a topological insulator for a certain range of composition x. ARPES experiments (73) have shown topologically nontrivial surface states in this system. However, the surface states in Bi_xSb_{1-x} are complicated and cannot be described by a simple model Hamiltonian.

The topological band theory is valid only for noninteracting systems in the absence of disorder. The topological field theory is a more general theory that describes the electromagnetic response of the topological insulator (23). Qi et al. (23) found that the electromagnetic response of three-dimensional topological insulators is described by the Maxwell equations with an added topological term proportional to $\mathbf{E}\cdot\mathbf{B}$. This exact modification (23) had been proposed earlier in the context of high-energy physics (74), as a modification to conventional electrodynamics due to the presence of the Peccei-Quinn axion field (75). In this approach (23), the $\mathbb{Z}_2$ topological invariant from topological band theory corresponds to a quantized emergent axion angle θ that is constrained by TR invariance to take only two values, 0 (the trivial insulator) or π (the topological insulator). The equivalence between the two definitions has recently been proven (76). Several unique experiments based on axion electrodynamics in three-dimensional topological insulators have been proposed: a topological Kerr and Faraday effect (23, 77–79), a topological magneto-electric effect (23), and an image magnetic monopole effect (80). Efforts toward the discovery of these exotic phenomena, as well as intensive searches for new three-dimensional topological insulator materials, are ongoing.

7. CONCLUSION AND OUTLOOK

This review covers our current theoretical understanding of the QSH state, with an emphasis on the theoretical prediction of the QSH state in the HgTe/CdTe QW system and its experimental realization in that particular material. We first discuss a phenomenological description of the QSH state in terms of two copies of the QH state for opposite spins and related by TR. As a consequence of this phenomenological description, we introduce the concept of helical edge state in terms of two copies of chiral QH edge states for opposite spins and related by TR. We then explain the importance of spin-orbit coupling and the phenomenon of band inversion in the HgTe/CdTe system, which is key for the realization of the QSH effect. From simple symmetry arguments, we describe the main properties of the low-energy effective Hamiltonian (the BHZ model) for the QSH state in HgTe/CdTe QWs. We then review the experimental discovery of the QSH state in HgTe/CdTe QWs. The occurrence of band inversion is confirmed by the observation of a re-entrant QH effect in the presence of an external magnetic field, and transport measurements provide strong evidence for the existence of extended helical edge channels. We discuss the theory of the helical edge state in more detail, with an emphasis on the stability of the edge state to disorder and interactions. We also discuss the existence of

spin-charge-separated solitons in the bulk of a QSH system as a direct measurable consequence of the bulk $\mathbb{Z}_2$ topology. Finally, we briefly discuss the three-dimensional generalization of the QSH state, the three-dimensional topological insulator state.

Topological insulators in two and three dimensions have been the subjects of tremendous investigation over the past few years, from both a theoretical and experimental point of view. However, many questions remain to be answered. On the experimental side, several new material candidates for both two-dimensional (48, 81, 82) (QSH) and three-dimensional (83–85) topological insulators await experimental verification. Even for available materials (HgTe/CdTe for the QSH effect, Bi_2Se_3 and related compounds for the three-dimensional topological insulator), most measurements confirm the existence of the boundary states but do not probe their intrinsically topological properties (fractional charge for the QSH effect; topological Kerr/Faraday effect, topological magneto-electric effect, and monopole effect for the three-dimensional topological insulator). On the theoretical side, many avenues are open for further investigation. Perhaps one of the most interesting questions concerns whether one can find fractional topological insulator states in strongly correlated materials. This question can be interpreted in (at least) two ways. One can consider fractional states in the sense of spin-charge separation in Mott insulators, i.e., a topological insulator of deconfined, dynamical spinons (86, 87) that carry spin but no charge. Another definition of a fractional topological insulator (20, 88–90) is more analogous to the fractional QH state and corresponds to a state with fractional bulk topological quantum number and deconfined fractionally charged quasiparticles, for instance a fractional axion angle in the bulk of a three-dimensional topological insulator (89, 90). In any event, the prediction, discovery, and recent study of the quantum spin Hall effect and topological insulators have brought together in an unexpected fashion insights from fields as diverse as semiconductor physics, solid state physics, materials science, spintronics, quantum field theory, topological field theory, and particle physics. We look forward to the many exciting future developments that surely lie ahead.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank colleagues at the University of Würzburg, in particular L.W. Molenkamp, H. Buhmann, E.M. Hankiewicz, M. König, and C. Brüne, for extensive collaborations. We also thank E. Berg, B.A. Bernevig, A. Karch, E.A. Kim, C.X. Liu, Y. Oreg, X.L. Qi, Z. Wang, C.J. Wu, and H. Yao for collaborations and discussions. This work is supported by the Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering, under contract DE-AC02-76SF00515. J.M. is supported by a Stanford Graduate Fellowship. T.L.H. was supported in part by the National Science Foundation grant DMR 0758462 at the University of Illinois and by the Institute for Condensed Matter Theory.

LITERATURE CITED

1. Landau LD, Lifshitz EM. 1980. *Statistical Physics*. Oxford, UK: Pergamon
2. von Klitzing K, Dorda G, Pepper M. 1980. *Phys. Rev. Lett.* 45:494–97; doi:10.1103/PhysRevLett.45.494
3. Tsui DC, Stormer HL, Gossard AC. 1982. *Phys. Rev. Lett.* 48:1559–62; doi:10.1103/PhysRevLett.48.1559

4. Abrahams E, Anderson PW, Licciardello DC, Ramakrishnan TV. 1979. *Phys. Rev. Lett.* 42:673–76; doi:10.1103/PhysRevLett.42.673
5. Qi XL, Wu YS, Zhang SC. 2006. *Phys. Rev. B* 74:045125; doi:10.1103/PhysRevB.74.045125
6. Thouless DJ, Kohmoto M, Nightingale MP, den Nijs M. 1982. *Phys. Rev. Lett.* 49:405–8; doi:10.1103/PhysRevLett.49.405
7. Simon B. 1983. *Phys. Rev. Lett.* 51:2167–70; doi:10.1103/PhysRevLett.51.2167
8. Zhang SC, Hansson TH, Kivelson S. 1989. *Phys. Rev. Lett.* 62:82–85; doi:10.1103/PhysRevLett.62.82
9. Zhang SC. 1992. *Int. J. Mod. Phys. B* 6:25–58; doi:10.1142/S0217979292000037
10. Girvin SM, MacDonald AH. 1987. *Phys. Rev. Lett.* 58:1252–55
11. Ambegaokar V, Halperin BI, Nelson DR, Siggia ED. 1980. *Phys. Rev. B* 21:1806–26; doi:10.1103/PhysRevB.21.1806
12. Fisher MPA, Lee DH. 1989. *Phys. Rev. B* 39:2756–59; doi:10.1103/PhysRevB.39.2756
13. Bernevig BA, Hughes TL, Zhang SC. 2006. *Science* 314:1757–61; doi:10.1126/science.1133734
14. König M, Wiedmann S, Brüne C, Roth A, Buhmann H, et al. 2007. *Science* 318:766–70; doi:10.1126/science.1148047
15. Murakami S, Nagaosa N, Zhang SC. 2003. *Science* 301:1348–51; doi:10.1126/science.1087128
16. Sinova J, Culcer D, Niu Q, Sinitsyn NA, Jungwirth T, MacDonald AH. 2004. *Phys. Rev. Lett.* 92:126603; doi:10.1103/PhysRevLett.92.126603
17. Murakami S, Nagaosa N, Zhang SC. 2004. *Phys. Rev. B* 69:235206; doi:10.1103/PhysRevB.69.235206
18. Murakami S, Nagaosa N, Zhang SC. 2004. *Phys. Rev. Lett.* 93:156804; doi:10.1103/PhysRevLett.93.156804
19. Kane CL, Mele EJ. 2005. *Phys. Rev. Lett.* 95:226801; doi:10.1103/PhysRevLett.95.226801
20. Bernevig BA, Zhang SC. 2006. *Phys. Rev. Lett.* 96:106802; doi:10.1103/PhysRevLett.96.106802
21. Haldane FDM. 1988. *Phys. Rev. Lett.* 61:2015–18; doi:10.1103/PhysRevLett.61.2015
22. Kane CL, Mele EJ. 2005. *Phys. Rev. Lett.* 95:146802; doi:10.1103/PhysRevLett.95.146802
23. Qi XL, Hughes TL, Zhang SC. 2008. *Phys. Rev. B* 78:195424; doi:10.1103/PhysRevB.78.195424
24. Min H, Hill JE, Sinitsyn NA, Sahu BR, Kleinman L, MacDonald AH. 2006. *Phys. Rev. B* 74:165310; doi:10.1103/PhysRevB.74.165310
25. Yao Y, Ye F, Qi XL, Zhang SC, Fang Z. 2007. *Phys. Rev. B* 75:041401(R); doi:10.1103/PhysRevB.75.041401
26. König M, Buhmann H, Molenkamp LW, Hughes T, Liu CX, et al. 2008. *J. Phys. Soc. Jpn.* 77:031007; doi:10.1143/JPSJ.77.031007
27. Novik EG, Pfeuffer-Jeschke A, Jungwirth T, Latussek V, Becker CR, et al. 2005. *Phys. Rev. B* 72:035321; doi:10.1103/PhysRevB.72.035321
28. Pfeuffer-Jeschke A. 2000. *Bandstruktur und Landau-Niveaus quecksilberhaltiger II-VI Heterostrukturen.* PhD thesis. Univ. Würzburg, Germany
29. Rothe DG, Reinthaler RW, Liu CX, Molenkamp LW, Zhang SC, Hankiewicz EM. 2010. *New J. Phys.* 12:065012
30. Dai X, Hughes TL, Qi XL, Fang Z, Zhang SC. 2008. *Phys. Rev. B* 77:125319; doi:10.1103/PhysRevB.77.125319
31. Zhou B, Lu HZ, Chu RL, Shen SQ, Niu Q. 2008. *Phys. Rev. Lett.* 101:246807; doi:10.1103/PhysRevLett.101.246807
32. Datta S. 1995. *Electronic Transport in Mesoscopic Systems*. Cambridge, UK: Cambridge Univ. Press
33. Roth A, Brüne C, Buhmann H, Molenkamp LW, Maciejko J, et al. 2009. *Science* 325:294–97; doi:10.1126/science.1174736
34. Büttiker M. 2009. *Science* 325:278–79; doi:10.1126/science.1177157
35. Büttiker M. 1986. *Phys. Rev. Lett.* 57:1761–64; doi:10.1103/PhysRevLett.57.1761
36. Büttiker M. 1988. *Phys. Rev. B* 38:9375–89; doi:10.1103/PhysRevB.38.9375
37. Xu C, Moore JE. 2006. *Phys. Rev. B* 73:045322; doi:10.1103/PhysRevB.73.045322
38. Wu CJ, Bernevig BA, Zhang SC. 2006. *Phys. Rev. Lett.* 96:106401; doi:10.1103/PhysRevLett.96.106401
39. Moore JE, Balents L. 2007. *Phys. Rev. B* 75:121306(R); doi:10.1103/PhysRevB.75.121306
40. Roy R. 2009. *Phys. Rev. B* 79:195321; doi:10.1103/PhysRevB.79.195321

41. Li D, Shi J. 2009. *Phys. Rev. B* 79:241303(R); doi:10.1103/PhysRevB.79.241303
42. Ström A, Johannesson H, Japaridze GI. 2010. *Phys. Rev. Lett.* 104:256804
43. Maciejko J, Qi XL, Zhang SC. 2010. *Phys. Rev. B* 82:155310
44. Tkachov G, Hankiewicz EM. 2010. *Phys. Rev. Lett.* 104:166803; doi:10.1103/PhysRevLett.104.166803
45. Meidan D, Oreg Y. 2005. *Phys. Rev. B* 72:121312(R); doi:10.1103/PhysRevB.72.121312
46. Maciejko J, Liu CX, Oreg Y, Qi XL, Wu CJ, Zhang SC. 2009. *Phys. Rev. Lett.* 102:256803; doi:10.1103/PhysRevLett.102.256803
47. Hou CY, Kim EA, Chamon C. 2009. *Phys. Rev. Lett.* 102:076602; doi:10.1103/PhysRevLett.102.076602
48. Liu CX, Hughes TL, Qi XL, Wang K, Zhang SC. 2008. *Phys. Rev. Lett.* 100:236601; doi:10.1103/PhysRevLett.100.236601
49. Nozières P. 1974. *J. Low Temp. Phys.* 17:31–42; doi:10.1007/BF00654541
50. Ström A, Johannesson H. 2009. *Phys. Rev. Lett.* 102:096806; doi:10.1103/PhysRevLett.102.096806
51. Tanaka Y, Nagaosa N. 2009. *Phys. Rev. Lett.* 103:166403; doi:10.1103/PhysRevLett.103.166403
52. Law KT, Seng CY, Lee PA, Ng TK. 2009. Quantum dot in 2D topological insulator: the two-channel Kondo fixed point. Unpubl. manuscript; arXiv:0904.2262
53. Teo JCY, Kane CL. 2009. *Phys. Rev. B* 79:235321; doi:10.1103/PhysRevB.79.235321
54. Kharitonov M. 2010. Interaction-enhanced ferromagnetic insulating state of the edge of a two-dimensional topological insulator. Unpubl. manuscript; arXiv:1004.0194
55. Qi XL, Hughes TL, Zhang SC. 2008. *Nat. Phys.* 4:273–76; doi:10.1038/nphys913
56. Su WP, Schrieffer JR, Heeger AJ. 1979. *Phys. Rev. Lett.* 42:1698–701; doi:10.1103/PhysRevLett.42.1698
57. Jackiw R, Rebbi C. 1976. *Phys. Rev. D* 13:3398–409
58. Thouless DJ. 1983. *Phys. Rev. B* 27:6083–87; doi:10.1103/PhysRevB.27.6083
59. Qi XL, Zhang SC. 2008. *Phys. Rev. Lett.* 101:086802; doi:10.1103/PhysRevLett.101.086802
60. Ran Y, Vishwanath A, Lee DH. 2008. *Phys. Rev. Lett.* 101:086801; doi:10.1103/PhysRevLett.101.086801
61. Laughlin RB. 1981. *Phys. Rev. B* 23:5632–33; doi:10.1103/PhysRevB.23.5632
62. Lee DH, Zhang GM, Xiang T. 2007. *Phys. Rev. Lett.* 99:196805; doi:10.1103/PhysRevLett.99.196805
63. Weeks C, Rosenberg G, Seradjeh B, Franz M. 2007. *Nat. Phys.* 3:796–801; doi:10.1038/nphys730
64. Paranjape MB. 1985. *Phys. Rev. Lett.* 55:2390–3; doi:10.1103/PhysRevLett.55.2390
65. Zhang HJ, Liu CX, Qi XL, Dai X, Fang Z, Zhang SC. 2009. *Nat. Phys.* 5:438–42; doi:10.1038/nphys1270
66. Liu CX, Qi XL, Zhang HJ, Dai X, Fang Z, Zhang SC. 2010. *Phys. Rev. B* 82:045122
67. Xia Y, Wray L, Qian D, Hsieh D, Pal A, et al. 2009. *Nat. Phys.* 5:398–402; doi:10.1038/nphys1274
68. Chen YL, Analytis JG, Chu JH, Liu ZK, Mo SK, et al. 2009. *Science* 325:178–81; doi:10.1126/science.1173034
69. Hsieh D, Xia Y, Wray L, Qian D, Pal A, et al. 2009. *Science* 323:919–22; doi:10.1126/science.1167733
70. Fu L, Kane CL, Mele EJ. 2007. *Phys. Rev. Lett.* 98:106803; doi:10.1103/PhysRevLett.98.106803
71. Roy R. 2009. *Phys. Rev. B* 79:195322; doi:10.1103/PhysRevB.79.195322
72. Fu L, Kane CL. 2007. *Phys. Rev. B* 76:045302; doi:10.1103/PhysRevB.76.045302
73. Hsieh D, Qian D, Wray L, Xia Y, Hor YS, et al. 2008. *Nature* 452:970–74; doi:10.1038/nature06843
74. Wilczek F. 1987. *Phys. Rev. Lett.* 58:1799–802; doi:10.1103/PhysRevLett.58.1799
75. Peccei RD, Quinn HR. 1977. *Phys. Rev. Lett.* 38:1440–43; doi:10.1103/PhysRevLett.38.1440
76. Wang Z, Qi XL, Zhang SC. 2010. *New J. Phys.* 12:065007
77. Karch A. 2009. *Phys. Rev. Lett.* 103:171601; doi:10.1103/PhysRevLett.103.171601
78. Tse WK, MacDonald AH. 2010. *Phys. Rev. Lett.* 105:057401.
79. Maciejko J, Qi XL, Drew HD, Zhang SC. 2010. *Phys. Rev. Lett.* 105:166803
80. Qi XL, Li R, Zang J, Zhang SC. 2009. *Science* 323:1184–87; doi:10.1126/science.1167747
81. Murakami S. 2006. *Phys. Rev. Lett.* 97:236805; doi:10.1103/PhysRevLett.97.236805
82. Shitade A, Katsura H, Kuneš J, Qi XL, Zhang SC, Nagaosa N. 2009. *Phys. Rev. Lett.* 102:256403; doi:10.1103/PhysRevLett.102.256403
83. Lin H, Wray LA, Xia Y, Jia S, Cava RJ, et al. 2010. *Nat. Mater.* 9:546–49
84. Chadov S, Qi XL, Kübler J, Fecher GH, Felser C, Zhang SC. 2010. *Nat. Mater.* 9:541–45

85. Lin H, Wray LA, Xia Y, Xu SY, Jia S, et al. 2010. Single-Dirac-cone Z2 topological insulator phases in distorted Li2AgSb-class and related quantum critical Li-based spin-orbit compounds. Unpubl. manuscript; arXiv:1004.0999
86. Young MW, Lee SS, Kallin C. 2008. *Phys. Rev. B* 78:125316; doi:10.1103/PhysRevB.78.125316
87. Pesin DA, Balents L. 2010. *Nat. Phys.* 6:376–81
88. Levin M, Stern A. 2009. *Phys. Rev. Lett.* 103:196803; doi:10.1103/PhysRevLett.103.196803
89. Maciejko J, Qi XL, Karch A, Zhang SC. 2010. Fractional topological insulators in three dimensions. Unpubl. manuscript; arXiv:1004.3628
90. Swingle B, Barkeshli M, McGreevy J, Senthil T. 2010. Correlated topological insulators and the fractional magnetoelectric effect. Unpubl. manuscript; arXiv:1005.1076