Genome Biology

RESEARCH Open Access

Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities

Matthew N Bainbridge[1,2], Min Wang[1], Yuanqing Wu[1], Irene Newsham[1], Donna M Muzny[1], John L Jefferies[3], Thomas J Albert[4], Daniel L Burgess[4] and Richard A Gibbs[1*]

Abstract

Background: Enrichment of loci by DNA hybridization-capture, followed by high-throughput sequencing, is an important tool in modern genetics. Currently, the most common targets for enrichment are the protein coding exons represented by the consensus coding DNA sequence (CCDS). The CCDS, however, excludes many actual or computationally predicted coding exons present in other databases, such as RefSeq and Vega, and non-coding functional elements such as untranslated and regulatory regions. The number of variants per base pair (variant density) and our ability to interrogate regions outside of the CCDS regions is consequently less well understood.

Results: We examine capture sequence data from outside of the CCDS regions and find that extremes of GC content that are present in different subregions of the genome can reduce the local capture sequence coverage to less than 50% relative to the CCDS. This effect is due to biases inherent in both the Illumina and SOLiD sequencing platforms that are exacerbated by the capture process. Interestingly, for two subregion types, microRNA and predicted exons, the capture process yields higher than expected coverage when compared to whole genome sequencing. Lastly, we examine the variation present in non-CCDS regions and find that predicted exons, as well as exonic regions specific to RefSeq and Vega, show much higher variant densities than the CCDS.

Conclusions: We show that regions outside of the CCDS perform less efficiently in capture sequence experiments. Further, we show that the variant density in computationally predicted exons is more than 2.5-times higher than that observed in the CCDS.

Background

Single nucleotide variants (SNVs) and short indels can be discovered by hybridization-based targeted enrichment, followed by high-throughput DNA sequencing. This 'capture sequencing' can target the protein coding regions of the genome, the 'exome', and provide a cost-effective alternative to whole genome sequencing (WGS) [1-6]. Capture sequencing has now been applied to the identification of pathogenic variants in several disease models [7-16] and in population studies comparing phenotypically normal individuals [17].

DNA may be enriched by a number of methods [1,4,5,18]. Here, we perform liquid-phase hybridization using biotinylated, DNA-oligonucleotide probes with a typical length of 60 to 80 bp. The probes are incubated with fragmented genomic DNA, after ligation with sequencing-platform specific adapters. Subsequently, the desired regions are recovered via streptavidin-coated magnetic beads with affinity for the biotinylated oligonucleotide probes. This approach has allowed interrogation of a human exome, beginning with as little as 1 μg of total DNA and with just 3 Gbp of total raw sequence [6].

The consensus coding DNA sequence (CCDS) [19] exons have been used most frequently to guide the design of capture reagents because their gene models are robust and encompass only approximately 30 Mbp. The CCDS gene collection, however, is defined by conservative criteria and lacks many of the genes found in other sets, such as RefSeq [20]. Even when two gene

* Correspondence: agibbs@bcm.edu
[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
Full list of author information is available at the end of the article

collections share the same core genes, the underlying exon content of those genes may vary. Consequently, the optimal design choice for targeting protein coding regions is a matter of ongoing concern.

In addition, many regions related to gene function, such as transcription factor binding sites, enhancer sites and UTRs, exist outside of the coding exome. These elements may also contribute to disease pathogenicity and are desirable components of target probe sets. Understanding the expected variant density of these regions and our ability to sequence them are therefore important considerations for designing future capture experiments.

In order to expand the regions that can be effectively targeted in capture sequence experiments, we designed two new capture reagents, termed the VCR-set and REC-set (Figure 1). The VCR-set targets the microRNA (miRNA) [21], Vega [22], CCDS, and RefSeq gene models, including predicted genes within RefSeq, with a total target size of 42 Mbp. To evaluate its ability to capture non-conserved UTRs, this design includes 8 Mbp of randomly selected UTR exons (see Materials and methods). Our second capture design, the REC-set (regulome, exons, conserved elements) aims to capture a total genomic region of 52 Mbp. In addition to the CCDS, RefSeq and Vega exons, the REC-set targets conserved UTR elements [23,24], exons that have been predicted computationally [25,26], as well as the regulome, that is, regions believed to be involved in transcription factor-mediated gene regulation [27-29]. These targeted regions represent a wide range of GC contents (Figure 2).

The reagents developed here permitted us to determine the relative 'capture ability' of subregions of the genome, compared to the CCDS. This measure can be conflated by biases introduced through the sequencing platform and alignment algorithms used. To assess the specific effect of capture on enrichment, we compared the levels of sequence coverage over specific genomic loci to that of non-enriched, WGS at the same loci.

Results

In total, we aligned more than 54 Gbp of capture sequence data derived from seven separate libraries and five DNA samples to the human reference genome. The data were generated using both VCR-set and REC-set capture designs, and SOLiD (single-end) and Illumina (paired-end) sequencing platforms.

For the REC-set design, two libraries were constructed from DNA samples obtained from human blood, from individuals of Hispanic ethnicity (L721, L722) for Illumina sequencing. One SOLiD library was constructed utilizing DNA from a HapMap cell-line (NA12812). The VCR-set design was used to capture fragments of DNA derived from two human blood samples from individuals of European ethnicity (C45, C6) followed by Illumina

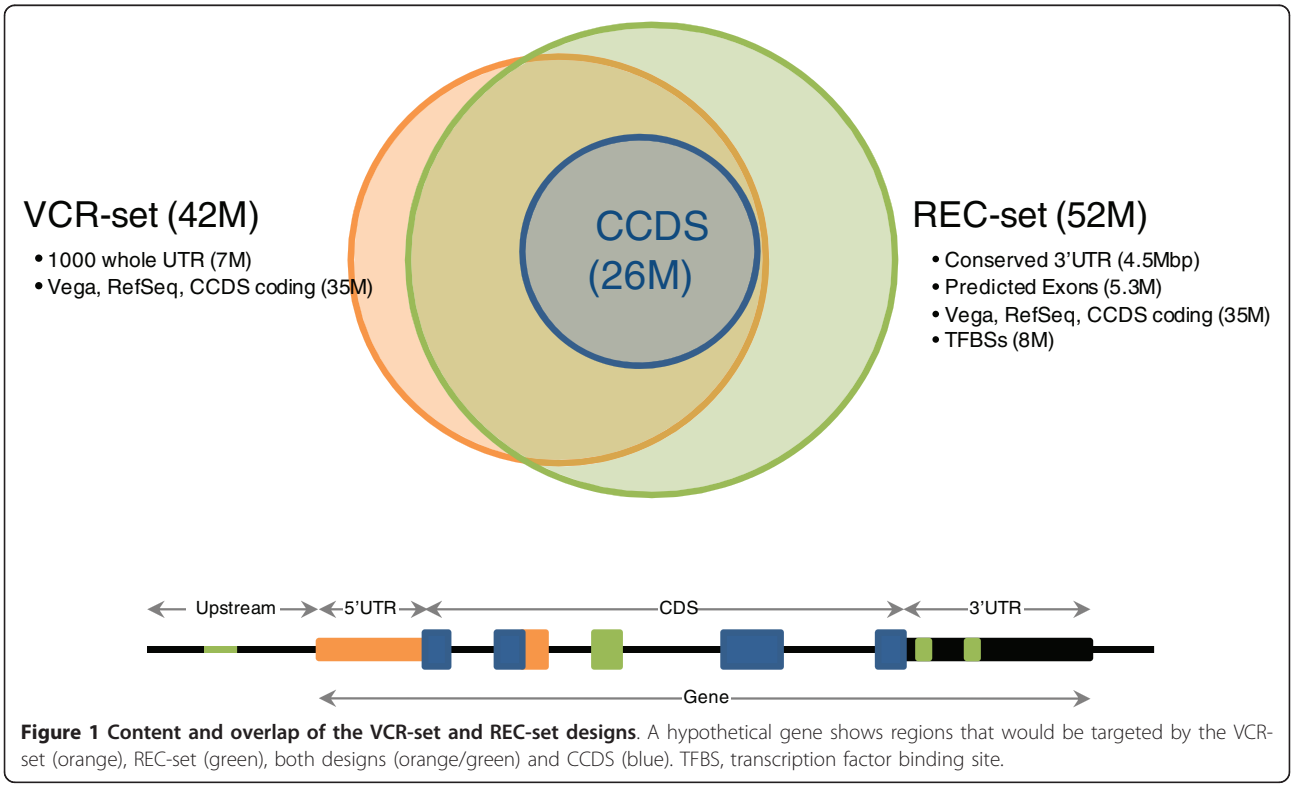

Figure 1 Content and overlap of the VCR-set and REC-set designs. A hypothetical gene shows regions that would be targeted by the VCR-set (orange), REC-set (green), both designs (orange/green) and CCDS (blue). TFBS, transcription factor binding site.

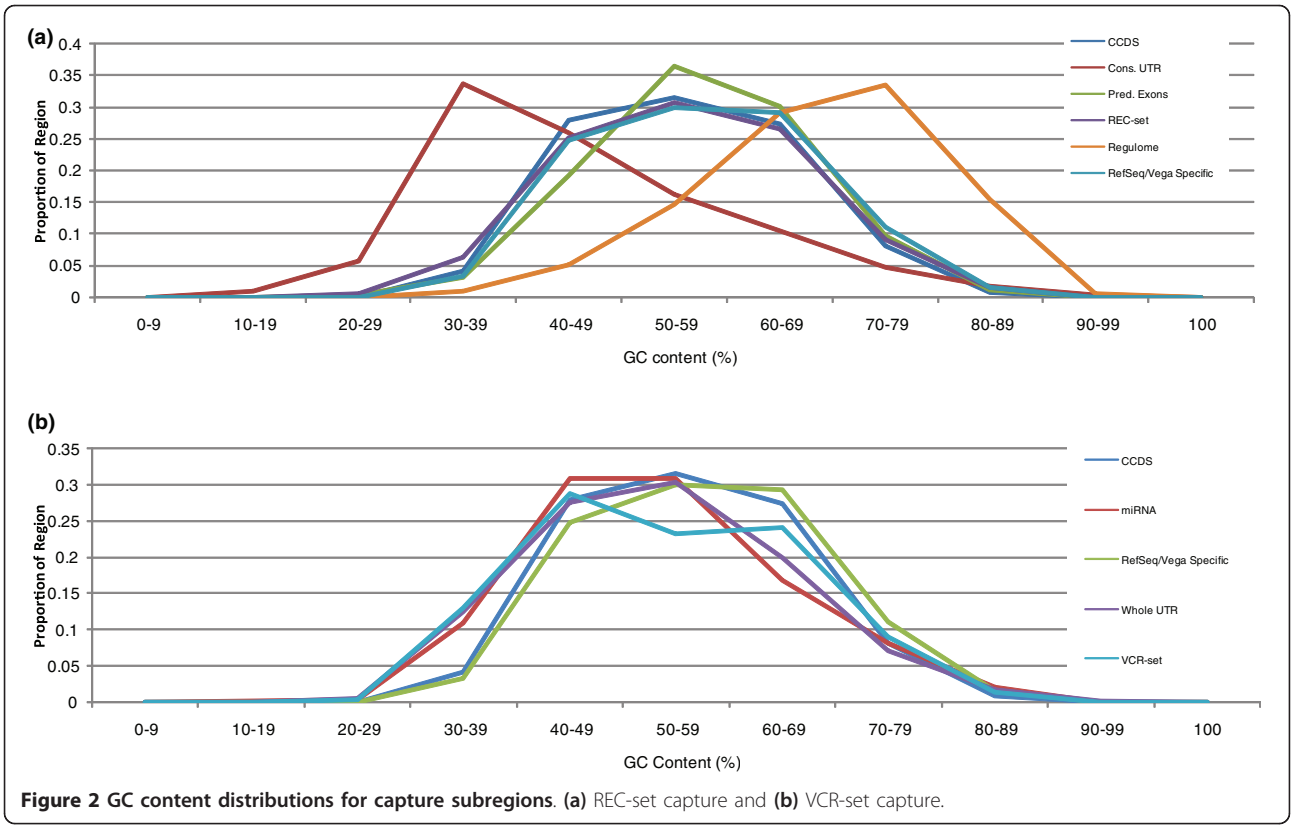

Figure 2 GC content distributions for capture subregions. **(a)** REC-set capture and **(b)** VCR-set capture.

sequencing, and two replicate libraries from the HapMap cell line DNA (NA12812) followed by SOLiD sequencing.

We had previously reported that sequence data from the Illumina platform routinely revealed a higher overall sequence capture enrichment yield than that found when using the SOLiD platform [6]. This was attributed to the relative ease of generating 'paired end' reads on the Illumina platform and the efficiency with which they could subsequently be precisely mapped to the genome, as well as differences in ligation strategies employed during library construction (Supplementary Table 1 in Additional file 1). Subsequent improvements in the library construction protocols for SOLiD capture sequencing have reduced the difference in the overall sequence capture enrichment yields between platforms (data not shown) [6]. To ensure meaningful comparison of data generated on the different sequencing platforms, we routinely ensured that the same number of targeted bases were covered at ≥10× for each experiment.

Capture efficiency and coverage

Genomic subregions were defined as groups of genomic segments with similar functional characteristics (UTRs, predicted exons, and so on; Figure 1). The 'capture-ability' of each subregion was defined as the average sequence coverage relative to the average coverage of the CCDS subregion. As can be seen in Figure 3a, the CCDS has approximately 10 to 15% higher average coverage than the REC-set target regions as a whole. Both the conserved UTR and regulome regions performed substantially worse than the CCDS, whereas predicted exons performed better than the CCDS. Most of the subregions in the VCR-set design performed within 10 to 20% of the CCDS (Figure 3b), the only exception being the non-conserved UTR subregion, which was substantially worse. Interestingly, the miRNA subregion performed slightly better than the CCDS when captured. For the majority of subregions, Illumina and SOLiD sequencing performed identically, with the exceptions of GC extremes (regulome, UTRs) and regions that are consistently represented by ~100 bp in the genome (for example, miRNA). Results differed by less than 1% from sample to sample (data not shown). Similar results were observed when considering the median level of coverage (Supplementary Figure 1 in Additional file 1), with the exception that the coverage performance of the REC-set and VCR-set as a whole was improved when compared to the CCDS. This is due to some CCDS targets having extremely high levels of coverage, which skews the mean coverage of these regions. Coverage differences between regions were found to be highly significant ($P < <$

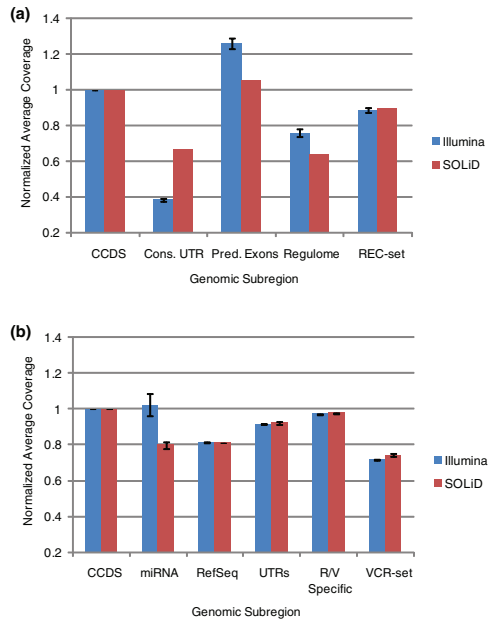

Figure 3 Coverage distributions for capture subregions. **(a)** Average coverage for subregions of REC-set design are shown as a proportion of the average coverage of the CCDS subregion. **(b)** Average coverage for subregions of VCR-set design are shown as a proportion of the average coverage of the CCDS subregion. 'R/V specific' refers to RefSeq/Vega exons not contained in the CCDS.

0.001) by the Mann-Whitney non-parametric test. Differences in coverage seem to be driven almost entirely by GC content. CCDS exons with very high GC content had coverage that was similar to regulatory regions with high GC content, whereas CCDS exons with very low GC content had depressed coverage similar to conserved UTRs (Supplementary Figure 2 in Additional file 1).

To determine whether the observed subregional differences in capture sequence data coverage resulted from variability in the capture efficiencies, or alternatively, by biases incurred during sequencing and alignment, we compared the capture SOLiD sequencing data to their equivalent regions in SOLiD WGS data [30]. First, we determined that no particular subregion was especially prone to mismapping by artificially generating reads from these regions and mapping them back to the genome and by comparing the mapping scores (a measure of the ratio of the best to the second best alignment score) of real data aligned to each subregion (Supplementary Tables 2 and 3 in Additional file 1) and found that regions with very high variant densities (regulome, predicted exons) had mapping scores and mappabilities that were very similar to regions with the lowest variant densities (conserved UTRs); we conclude from this that the high observed variant densities in the predicted exons and regulome are likely not due to mismapping of reads.

In general, the relative coverage patterns observed in WGS were similar to those from captured material (Figure 4a); however, two regions (miRNA and predicted exons) performed better than expected when captured (a positive value in Figure 4b), whereas both UTR regions performed substantially worse than WGS. These data show that some biases in recovery of sequence data from some genomic regions were incurred during sequencing, and not during the earlier capture phase.

Variant density in subregions

We examined the density of SNV sites in different capture subregions. All data were filtered to retain reads with high mapping qualities and regions with 10× or higher sequence read coverage (see Materials and methods; Supplementary Table 3 in Additional file 1). There were differences in the discovery rate with different platforms (Illumina approximately 1/1,500 bp in the CCDS exome versus approximately 1/1,700 bp for SOLiD; Supplementary Table 4a, b in Additional file 1). Both values were similar to the variant densities observed in other exon studies (for example, Thousand Genomes Pilot Three [17]), but were considerably lower than those previously reported for the whole genome (approximately 1/1,000 bp) [31-33].

The evolutionarily conserved UTR portion of the REC-set design harbored 10 to 25% fewer (Figure 5a) variants (1/2,300 bp SOLiD, 1/1,625 bp Illumina) than the CCDS exome, in stark contrast to the non-conserved UTR portion of the VCR-set design (Figure 5b; Supplementary Table 4b in Additional file 1), which showed an 80 to 100% increase (1/925 bp SOLiD, 1/750 bp Illumina). Thus, there was an approximately 2.5-fold differential in variant density between conserved and non-conserved regions of the UTR.

Surprisingly, the predicted exons and regulatory regions in the REC-set design exhibited more than two times the variant density observed in the CCDS exome, a value higher than the average rate of the whole genome (1/600 bp to 1/800 bp). This suggested that these regions were either more tolerant to variation, or that these regions have increased mutation rates compared to the whole genome. Increasing or decreasing the stringency of the variant calling parameters had little effect on either the absolute variant density or the density relative to the CCDS exome (data not shown).

To confirm that this observation was not an artifact of allele-bias during the capture process, we compared these results to those obtained from two WGS SOLiD data sets [30,31] and filtered variants for predicted exons and CCDS target regions. As expected, the African genome that was previously sequenced showed a higher variant density throughout the genome (1/900 bp) than the Caucasian genome (1/1,061 bp) as well as

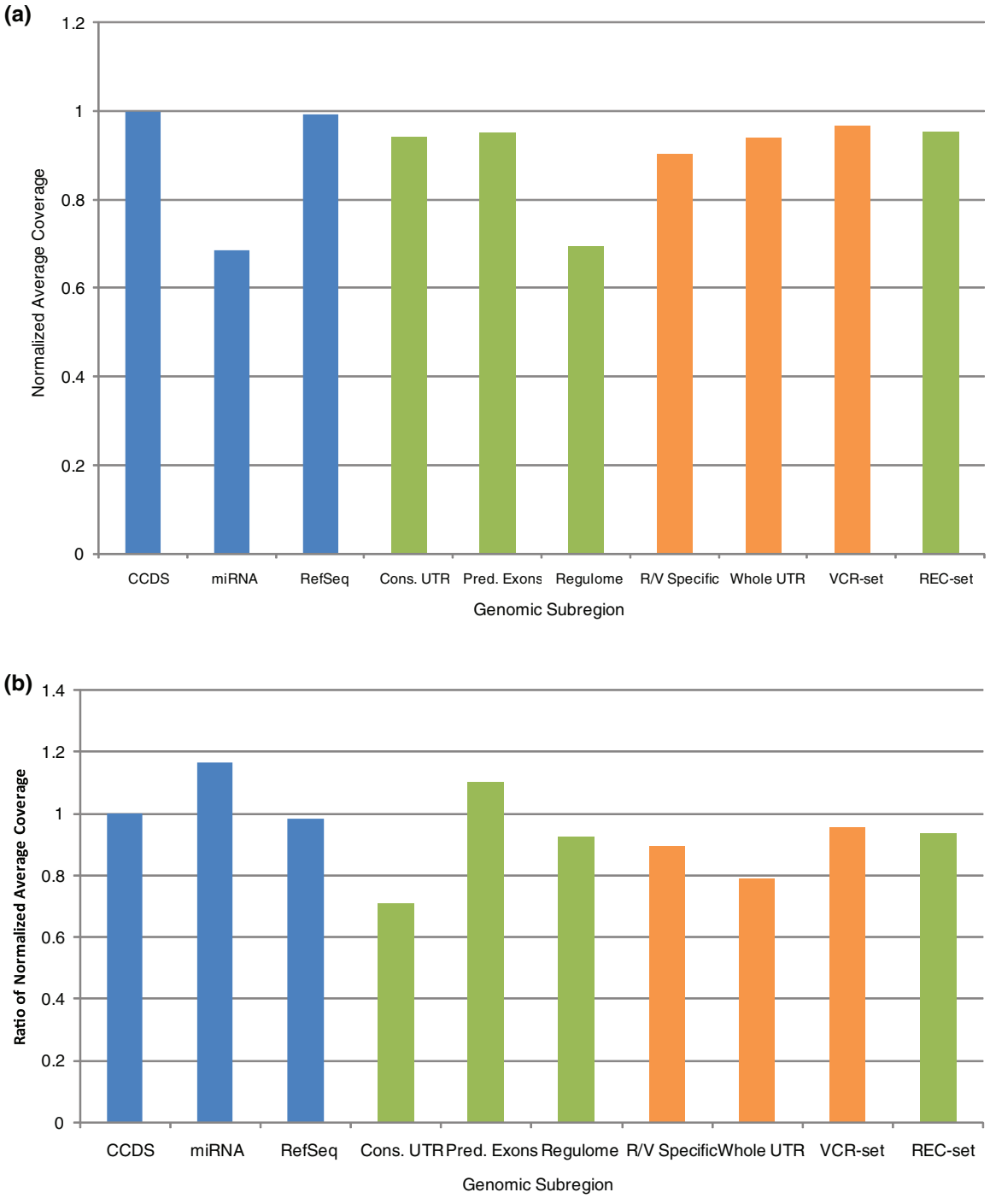

Figure 4 Normalized coverage distributions. **(a)** Coverage of genomic subregions, relative to the CCDS, after whole genome SOLiD sequencing. Green, regions specific to REC-set; orange, regions specific to VCR-set; blue, shared regions. 'R/V specific' refers to RefSeq/Vega exons not contained in the CCDS. **(b)** Proportional difference in relative coverage between capture-sequencing and WGS shows both enrichment (values > 1) and depletion (values < 1) of certain genomic subregions after capture. Green, regions specific to REC-set; orange, regions specific to VCR-set; blue shared regions.

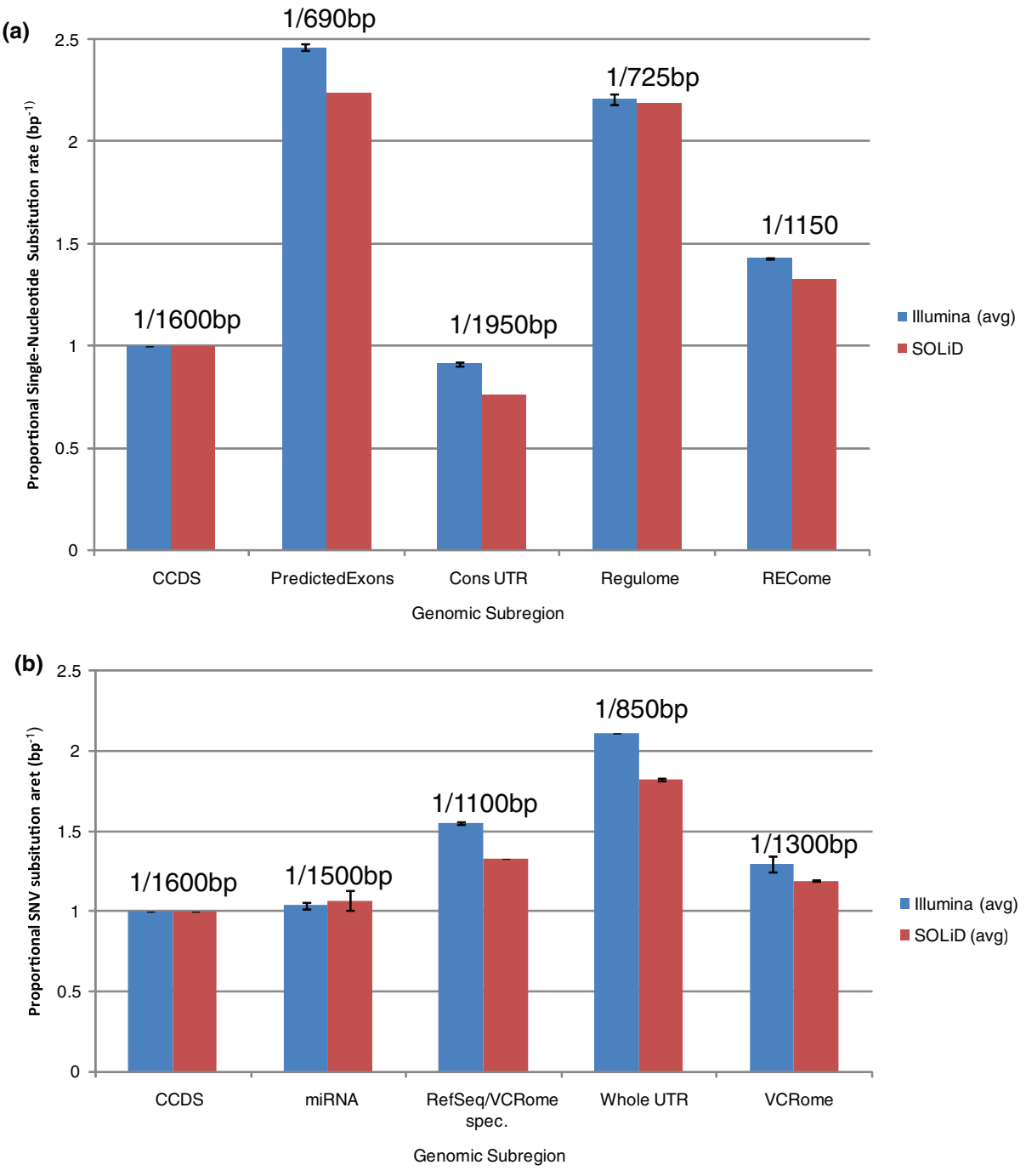

Figure 5 Single nucleotide variant densities. **(a)** Number of SNV substitutions per base pair, of REC-set subregions, as a proportion of the SNV rate of the CCDS subregion. The absolute average value from SOLiD and Illumina sequencing is given above the data point. **(b)** Number of SNV substitutions per base pair, of VCR-set subregions, as a proportion of the SNV rate of the CCDS subregion. The absolute average value from SOLiD and Illumina sequencing is given above the data point. 'R/V specific' refers to RefSeq/Vega exons not contained in the CCDS.

in each subregion examined here (Table 4c in Additional file 1). The relative variant densities, however, when normalized to the CCDS exome, were approximately the same in both genomes. The observed high density of variants in the predicted exons was even more pronounced in these data, with predicted exon regions (1/714 bp) having approximately 2.5× the mutation rate of the CCDS (1/1,808 bp), and a 30 to 40% increase over the genome as a whole. To ensure that these results were not an artifact of high-throughput sequencing, we examined the variant density of these regions in HuRef [33]. The variant density in the CCDS region of HuRef was slightly depressed compared to the other two WGS datasets, but the relative variant density for each region was similar or more pronounced (Supplementary Table 4c in Additional file 1).

We examined the mutation spectrum of the observed variants in different subregions (Additional file 2). Interestingly, we found the transition:transversion ratio in the CCDS region to be 3:1, in both capture and WGS datasets, whereas across the whole genome the rate was approximately 2:1 [34]. This value was also higher than that seen in the regulome of 1.6:1. The mutation spectrum was significantly different for the regulome compared to the CCDS exome; the number of C→T and G→A mutations, as a proportion of the total number of mutations, was significantly repressed in the regulome compared to the CCDS, despite having a higher GC content and a higher proportion of CpG dinucleotides, which are known to be prone to mutation [35]. Interestingly, predicted exon subregions showed intermediate levels of all mutation types when compared to the CCDS and regulome regions. This implies that the mutation spectrum alone cannot account for the observed variant density.

Lastly, we hypothesized that predicted exons may have variant density properties identical to that of introns. Introns are thought to have a higher variant density than the whole genome because they are frequently transcribed [36]. Because no specifically intronic regions were captured by our designs, we used the WGS data and found that the intronic variant density (approximately 1/850) is slightly higher than that of the whole genome (approximately 1/1,000), but still significantly lower (*P*-value of approximately 0.0001) than that of predicted exons (approximately 1/700) (Table 4c in Additional file 1). It remains possible that the GC content of the predicted exons makes the variant density higher than that seen in the remainder of the intron.

The high variant density we observed in the predicted exons led us to examine the evolutionary conservation of these regions relative to the introns and CCDS exons. As expected, the CCDS exons were highly conserved relative to the intronic regions. Although the predicted exons mimicked the intronic regions by having a large proportion of bases with neutral evolution scores, these regions had more bases with both high and low conservation when compared to the introns (Figure 6). We next examined the minor allele frequency distribution, using data from the Thousand Genomes Project [17], of

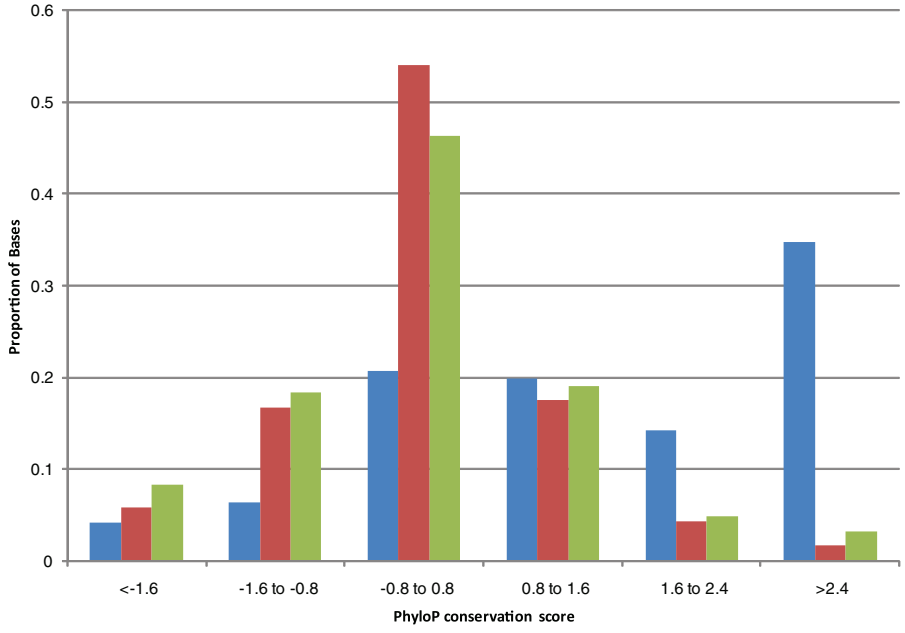

Figure 6 Distribution of phyloP scores across the CCDS (blue), intronic (red) and predicted exons (green).

variants in the intronic, CCDS exome and predicted exon regions for HuRef (Figure 7). Minor allele frequency distributions for CCDS and predicted exons in capture data were similar (data not shown). Although 12% of the intronic variants and only 9% of the CCDS variants were private (unseen in public databases), fully approximately 16% of the predicted exon variants were not found in data from the Thousand Genomes Project. This situation was reversed for fixed variants, with predicted exons having the smallest proportion (approximately 4%) compared to CCDS variants (approximately 6%).

Discussion

In this study we show the efficacy of DNA capture sequencing and interrogation of variants in biologically important loci outside of the CCDS exome. These regions almost uniformly demonstrated decreased capture ability, as measured by average target coverage, when compared to the CCDS regions. Overall, both Illumina and SOLiD sequencing platforms showed similar biases in coverage of genomic subregions when measured relative to the CCDS. Importantly, capture ability appeared to be confounded by biases introduced by the sequencing technology and correlated with GC content of the target sequence, a known factor in short-read sequencing [37,38]. Particularly, conserved UTR regions, which are approximately 30% GC, and regulatory regions, which are approximately 70% GC, had approximately half of the sequence depth of coverage as the CCDS regions, approximately 50% GC. When compared to WGS (non-capture) data the same general biases were evident. However, the act of capturing the targeted regions seems to exacerbate the coverage bias by an additional 5 to 10%. The exceptions to this are the predicted exons and microRNA, where the coverage was higher than expected and the UTR regions where the coverage was as much as 25% lower than expected from the WGS data. This effect may be due to steric hindrance of probe-target binding introduced by secondary structure present in the UTR regions. These results imply that naively capturing biologically relevant loci other than the CCDS will require 20 to 40% more sequencing data to be generated than expected from the CCDS. It may be possible, however, to alter the capture reagent, perhaps by increasing the representation of some probes, in order to compensate for the empirically measured coverage biases and thus help normalize the coverage when capturing CCDS and other elements.

To our knowledge, this is the first targeted-sequence capture study of a genome-wide, diverse set of biologically important elements, allowing the investigation of variant densities in functionally relevant loci that have been hitherto undetected at a fraction of the cost of

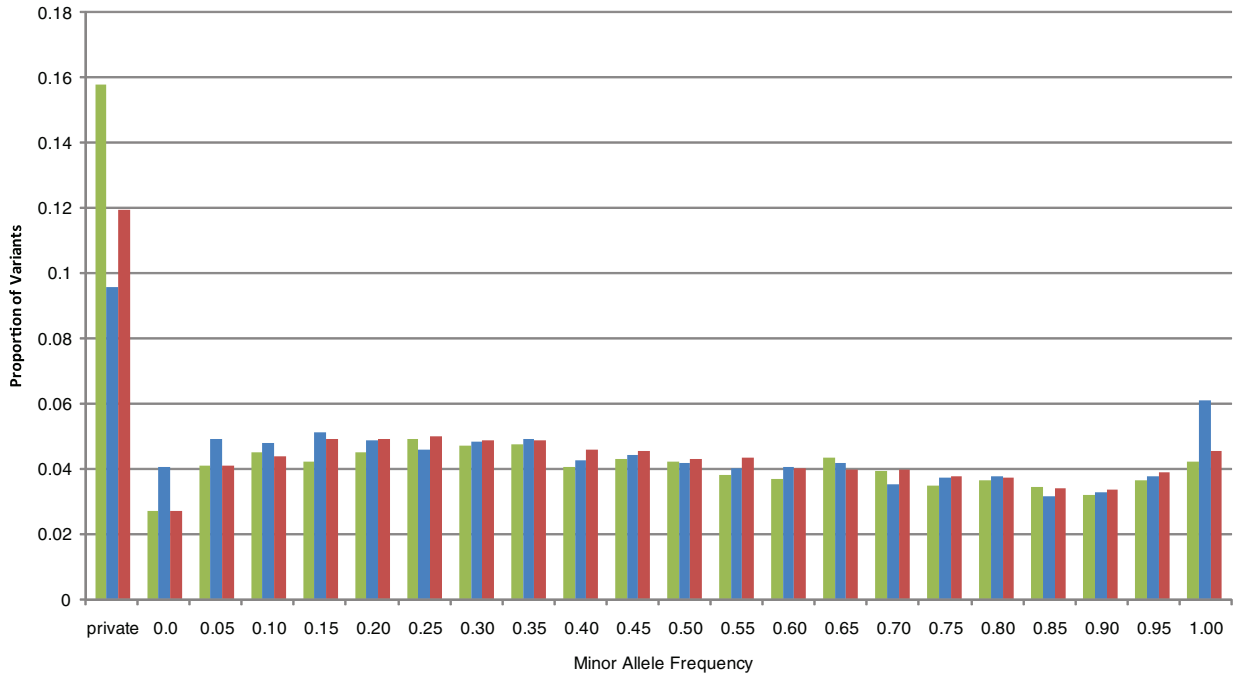

Figure 7 Minor allele frequency distributions for variants in HuRef subregions: predicted exons (green), CCDS exons (blue) and introns (red). 'Private' indicates the variant was not found in the Thousand Genomes Project.

whole genome sequencing. Using both Illumina and SOLiD sequencing, we demonstrate the ability to find variants across a significantly larger target region than the CCDS. As capture sequencing enables high levels of sequence coverage, we were able to discover rare (private) variants in each sample, using similar amounts of data to that used by low-coverage, whole-genome techniques that are better suited for common variant discovery.

Illumina sequencing consistently showed higher variant densities than SOLiD sequencing. This discrepancy is likely due to differences in variant filtering parameters used for the two different sequencing types. However, it may also reflect the inherently higher accuracy of SOLiD sequencing [37]. Importantly, when measured relative to the CCDS variant density, different subregions showed remarkably similar variant densities for both sequencing platforms. Variant densities, however, were found to vary in different subregions of the genome, likely due to evolutionary conservation and base composition of these regions. The evolutionarily conserved CCDS exome and UTR regions showed variant densities of 1/1,600 to 1/1,850 bp, considerably less than the whole genome rate of 1/1,000 bp, which presumably reflects the result of purifying selection acting to remove deleterious variants. Exons specific to RefSeq, which are not in the CCDS, showed intermediate levels of variant density, 1/1,200 bp. This is likely because these loci are less essential to the organism, and mutations in these regions are less likely to be deleterious. Unlike the coding regions, the regulome showed a variant density higher than the whole genome. While this is likely due to the GC content of the regulome, we found that C→T and G→A mutations were underrepresented as a portion of all variants when compared to the CCDS. This is significant because 5-methyl-cytosine bases in CpG dinucleotides, which are over-represented in regulatory regions, are prone to spontaneous deamination to uracil and subsequent repair to thymine [39]. This would indicate there is strong selective pressure to maintain cytosine and guanine representation in the regulome compared to the CCDS exome.

Of all the regions interrogated, the predicted exons showed the highest variant density, 1/660 bp. Although these exons have a higher GC content than the CCDS, it is considerably lower than the regulome, indicating that the increased mutability of GC-rich sequence content cannot fully account for the variant density. However, we observed that the intronic variant density in WGS studies was also considerably higher than that of the whole genome. It has been reported that transcribed regions have higher variant densities than non-transcribed regions [40,41] and we surmise, therefore, that the observed variant density is a combination of these regions being actively transcribed and their high GC content. As expected from the high variant density, predicted exon regions showed a slightly higher proportion of bases with faster than neutral evolution rates than when compared to intronic regions. Unexpectedly, predicted exons also showed a slightly higher proportion of conserved bases when compared to intronic regions.

The 'exonization' of intronic elements is well documented [42-45] and computationally predicted exons have been detected in mature mRNA from RNA-seq experiments [46]. In this work we interrogated predicted exons that are flanked by canonical splice-sites and exist within known CCDS genes and thus are good candidates for inclusion in mature RNA and subsequent translation. Exons are thought to be protected from mutation [47] and the higher mutation rates in predicted-exons may then be a source of evolutionary diversity.

Conclusions

This work has important implications for the large number of CCDS-based exome-capture experiments currently being reported. Specifically, caution should be used when extrapolating CCDS results to the entire human exome. Regions outside of the CCDS are more difficult to sequence, map and capture and require more raw sequence data than otherwise expected. Further, studies that seek to characterize human coding variation across a large number of individuals should use a diverse set of gene models to better measure and understand variation in less conserved coding elements. Consideration of non-CCDS regions in general will complicate sequence-based genome-wide disease studies due to the wider range of variant densities they exhibit, and will necessitate innovative bioinformatic data filtering strategies.

Materials and methods

DNA

DNA was obtained from the Corriel biorepository (catalog id GM12812). DNA was obtained from individuals under written informed consent for participation in the study. The study was approved by the institutional review board at Baylor College of Medicine and was conducted in accordance with the Helsinki declaration.

Target regions and probe design

All annotations, except for miRNA, were downloaded from UCSC Genome Browser [48] (hg18) on 1 October 2009. Coding regions: exons for the CCDS, RefSeq and Vega gene sets were all obtained in their entirety and non-coding regions were removed internally. Conserved UTRs: UTR regions were selected from the RefSeq and Vega gene sets. A region was considered

conserved if it had an LOD score ≥100 as determined by the phastCons [24] package using 17 vertebrate genomes (17-way most conserved track). miRNA: these annotations were obtained from miRNA base v13. Predicted exons: these annotations were obtained from Contrast and GenScan. Only predicted exons that occurred within the introns of known genes were used. Regulatory regions: these regions were obtained from the ORegAnno track and the Hudson Alpha transcription factor binding site (CHiP-seq) tracks. Only ORegAnno annotations that were < 50 bp were considered so as to remove non-transcription factor binding site regulatory regions and limit the total target size. CHiP-seq sites were only considered if they had an enrichment score of 300 or greater. Solution capture probes were designed and produced by Roche NimbleGen (Madison WI, USA) as previously described [6]. Nonconserved UTR: approximately 1,500 exons were randomly selected from 5' and 3' UTRs of Vega genes without any consideration for their conservation; however, they were always either the first or last exon in the annotated gene. RefSeq/Vega-specific regions: these regions are derived by algorithmically subtracting the CCDS regions from the combined Vega/RefSeq regions. For the purposes of this paper, derived regions < 50 bp were not considered as small regions, and regions at the edges of targets show lower coverage, generally.

Target regions are available as BED files in Additional files 3 and 4.

Library preparation

Precapture libraries for SOLiD (2 μg) were hybridized in solution according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos TrTA-A and SOLiD-B replaced oligos PE-HE1 and PE-HE2 and post-capture ligation-mediated PCR was performed using 12 cycles. Capture libraries were quantified using PicoGreen (catalog number P7589) and their size distribution analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500 (catalog number 5067-1506). Capture efficiency was evaluated by performing a quantitative PCR-based SYBR Green assay (Applied Biosystems, Foster City CA, USA Inc.; catalog number 4368708) with built-in controls (RUNX2, PRKG1, SMG1, and NLK). Capture library enrichment was estimated at seven to nine cycles over background by quantitative PCR. Captured libraries were further processed for sequencing, with approximately 6 to 12 Gbs of sequence generated per capture library on either SOLiD V3 or V4 instruments (Applied Biosystems, Inc.). A complete capture protocol can be found on the Baylor Human Genome Website [49]. Illumina library preparation was conducted as previously described [6].

Sequence data generation, alignment and variant calling

SOLiD data were aligned to the human genome (hg18) with BFast [50] and Illumina data with BWA [51]. Variants were filtered for quality as previously described [6]. Briefly, read qualities were recalibrated with GATK and a minimum quality score of 30 was required; also, the variant must have been present in at least 15% of the reads that cover the position. In addition, prior to variant calling reads with low (< 11) mapping qualities (a value based on the ratio of the best alignment score to the second best alignment score) were removed. This typically eliminates approximately 5 to 10% of the aligned reads. Sequence data were produced from either SOLiDv3 or Illumina GAII sequencing machines and are available from the Sequence Read Archive [52] with accession [SRP004501.1].

Additional material

Additional file 1: Supplementary data and statistics. Experimental design, capture statistics, regional description statistics, as well as whole genome statistics.

Additional file 2: Mutation spectrum. The mutation spectrum, transition:transversion ratio of discovered variants ordered by subregion.

Additional file 3: REC-set targets. Targeted regions in the REC-set.

Additional file 4: VCR-set targets. Targeted regions in the VCR-set.

Abbreviations

bp: base pair; CCDS: consensus coding DNA sequence; Gbp: Giga-base pair; Mbp: Mega-base pair; miRNA: microRNA; SNV: single nucleotide variant; UTR: untranslated region; WGS: whole genome sequencing.

Acknowledgements

The authors would like to thank Svasti Haricharan for editing the manuscript. This project was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) PGSD scholarship and by Award Number U54HG003273 from the National Human Genome Research Institute.

Author details

[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [2]Department of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [3]Department of Pediatrics-Cardiology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [4]Roche NimbleGen, Inc., 504 S. Rosa Road, Madison, WI 53719, USA.

Authors' contributions

MNB aided in experiment design, analysis and manuscript preparation. MW, YQW and DM conducted capture hybridization and sequencing. JLJ provided DNA samples and helped in data interpretation. DLB, TA and RAG participated in experimental design, capture design and drafting the manuscript. All authors have approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 23 November 2010 Revised: 16 January 2011
Accepted: 25 July 2011 Published: 25 July 2011

References

1. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA: **Direct selection of human genomic loci by microarray hybridization.** *Nat Methods* 2007, **4**:903-905.
2. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR: **Genome-wide *in situ* exon capture for selective resequencing.** *Nat Genet* 2007, **39**:1522-1527.
3. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461**:272-276.
4. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: **Microarray-based genomic selection for high-throughput resequencing.** *Nat Methods* 2007, **4**:907-909.
5. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J: **Multiplex amplification of large sets of human exons.** *Nat Methods* 2007, **4**:931-936.
6. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA, Jeddeloh JA, Muzny D, Albert TJ, Gibbs RA: **Whole exome capture in solution with 3 Gbp of data.** *Genome Biol* 2010, **11**:R62.
7. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: **Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.** *Proc Natl Acad Sci USA* 2009, **106**:19096-19101.
8. Otto EA, Hurd TW, Airik R, Chaki M, Zhou W, Stoetzel C, Patil SB, Levy S, Ghosh AK, Murga-Zamalloa CA, van Reeuwijk J, Letteboer SJ, Sang L, Giles RH, Liu Q, Coene KL, Estrada-Cuzcano A, Collin RW, McLaughlin HM, Held S, Kasanuki JM, Ramaswami G, Conte J, Lopez I, Washburn J, Macdonald J, Hu J, Yamashita Y, Maher ER, Guay-Woodford LM, *et al*: **Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy.** *Nat Genet* 2010, **42**:840-850.
9. Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, Kaymakcalan H, Barak T, Bakircioglu M, Yasuno K, Ho W, Sanders S, Zhu Y, Yilmaz S, Dincer A, Johnson MH, Bronen RA, Kocer N, Per H, Mane S, Pamir MN, Yalcinkaya C, Kumandas S, Topcu M, Ozmen M, Sestan N, *et al*: **Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations.** *Nature* 2010, **467**:207-210.
10. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42**:30-35.
11. Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK, Chong K, Mullikin JC, Biesecker LG: **Massively parallel sequencing of exons on the × chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate.** *Am J Hum Genet* 2010, **86**:743-748.
12. Rehman AU, Morell RJ, Belyantseva IA, Khan SY, Boger ET, Shahzad M, Ahmed ZM, Riazuddin S, Khan SN, Riazuddin S, Friedman TB: **Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79.** *Am J Hum Genet* 2010, **86**:378-388.
13. Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Kohler S, Jager M, Grunhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN: **Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome.** *Nat Genet* 2010, **42**:827-829.
14. Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M: **Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82.** *Am J Hum Genet* 2010, **87**:90-94.
15. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, Roepman R, Knoers NV, Veltman JA, Brunner HG: **Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome.** *Am J Hum Genet* 2010, **87**:418-423.
16. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, Veltman JA: ***De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome.** *Nat Genet* 2010, **42**:483-485.
17. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061-1073.
18. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nat Biotechnol* 2009, **27**:182-189.
19. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, *et al*: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res* 2009, **19**:1316-1323.
20. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, , **35(Database issue)**: D61-D65.
21. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, , **36(Database issue)**: D154-D158.
22. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database.** *Nucleic Acids Res* 2008, , **36(Database issue)**: D753-D760.
23. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, , **37(Database issue)**: D755-D761.
24. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15**:1034-1050.
25. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.
26. Gross SS, Do CB, Sirota M, Batzoglou S: **CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction.** *Genome Biol* 2007, **8**:R269.
27. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ: **ORegAnno: an open-access community-driven resource for regulatory annotation.** *Nucleic Acids Res* 2008, , **36(Database issue)**: D107-D113.
28. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJ: **ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.** *Bioinformatics* 2006, **22**:637-640.
29. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462**:315-322.
30. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, *et al*: **Complete Khoisan and Bantu genomes from southern Africa.** *Nature* 2010, **463**:943-947.
31. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F,

Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med* 2010, **362**:1181-1191.
32. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872-876.
33. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, *et al*: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5**:e254.
34. **The International HapMap Project..** *Nature* **426**:789-796.
35. Hwang DG, Green P: **Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution.** *Proc Natl Acad Sci USA* 2004, **101**:13994-14001.
36. Green P, Ewing B, Miller W, Thomas PJ, Green ED: **Transcription-associated mutational asymmetry in mammalian evolution.** *Nat Genet* 2003, **33**:514-517.
37. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, *et al*: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, **19**:1527-1541.
38. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**:e105.
39. Shen JC, Rideout WM, Jones PA: **The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA.** *Nucleic Acids Res* 1994, **22**:972-976.
40. Datta A, Jinks-Robertson S: **Association of increased spontaneous mutation rates with high levels of transcription in yeast.** *Science* 1995, **268**:1616-1619.
41. Bachl J, Carlson C, Gray-Schopfer V, Dessing M, Olsson C: **Increased transcription levels induce higher mutation rates in a hypermutating cell line.** *J Immunol* 2001, **166**:5051-5057.
42. Lev-Maor G, Sorek R, Shomron N, Ast G: **The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons.** *Science* 2003, **300**:1288-1291.
43. Sorek R: **The birth of new exons: mechanisms and evolutionary consequences.** *Rna* 2007, **13**:1603-1608.
44. Makalowski W, Mitchell GA, Labuda D: **Alu sequences in the coding regions of mRNA: a source of protein variability.** *Trends Genet* 1994, **10**:188-193.
45. Maia RM, Valente V, Cunha MA, Sousa JF, Araujo DD, Silva WA Jr, Zago MA, Dias-Neto E, Souza SJ, Simpson AJ, Monesi N, Ramos RG, Espreafico EM, Paco-Larson ML: **Identification of unannotated exons of low abundance transcripts in Drosophila melanogaster and cloning of a new serine protease gene upregulated upon injury.** *BMC Genomics* 2007, **8**:249.
46. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470-476.
47. Kaplan CD: **Revealing the hidden relationship between nucleosomes and splicing.** *Cell Cycle* 2009, **8**:3633-3634.
48. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ: **The UCSC genome browser database: update 2007.** *Nucleic Acids Res* 2007, , **35(Database issue)**: D668-D673.
49. **Nimblegen Capture Protocol for SOLiD Platform..** [http://www.hgsc.bcm.tmc.edu/cascade-tech-solid_capture_protocol-st.hgsc?pageLocation=solid_capture_protocol].
50. Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing.** *PLoS One* 2009, **4**:e7767.
51. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
52. **Sequence Read Archive..** [http://www.ncbi.nlm.nih.gov/Traces/sra].

doi:10.1186/gb-2011-12-7-r68
Cite this article as: Bainbridge *et al.*: Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biology* 2011 **12**:R68.