Lessons learned from sequencing under a linkage peak in families with vesicoureteric reflux

Heather J. Cordell

Institute of Genetic Medicine
Newcastle University, UK

heather.cordell@ncl.ac.uk

Affected sib pair study in vesicoureteric reflux (VUR)

- Collaborative project dating from ∼2001
  - With Judith and Tim Goodship from Newcastle University
  - And collaborators from London (Adrian Woolf, Sue Malcom), Leeds (Sally Feather) and Slovenia (Rajko Kenda)
- VUR is a kidney disease characterised by abnormal movement of urine from the bladder into the ureter or kidneys
  - Nephropathy associated with VUR accounts for 10% of kidney failure requiring dialysis or transplantation
- Evidence for genetic/familial component
  - VUR present in 30-50% of siblings and 65% of offspring of affected individuals; 80% MZ twin concordance

UK VUR sib pair collection

- In 2001, we obtained Wellcome Trust funding to collect DNA from $\sim$ 300 UK affected sib pairs with VUR (plus parents)
  - Principal investigators: Adrian Woolf, Tim Goodship
  - Co-investigators: Sue Malcom, Sally Feather, Judith Goodship
  - Collaborator (statistical genetics): Heather Cordell
- Aim was to carry out a genome-wide linkage scan
  - Following on from prior genome-wide linkage scan in 7 large European VUR families (Feather et al. 2000)
    - Feather et al. (2000) had found significant evidence (p=0.0001) localising a genetic effect to chromosome 1
    - Evidence of heterogeneity, as 2 families not consistent with this genetic locus

VUR study

- UK VUR affected sib pair study
  - Designed as non-parametric linkage study
  - Originally aimed for $>$ 300 ASPs
    - Actually achieved 245 (165 families)
    - Supplemented by an additional 186 ASPs (149 families) from Slovenia
- WT funded the sample collection but not genotyping
  - In 2006, after an unsuccessful re-application to WT, MRC funding was obtained to carry out genotyping of affected children plus their parents

Genotyping platform

- In 2001 application, had planned to genotype using standard ABI microsatellite panel (300-400 markers, 10cM spacing)
  - Natural set for non-parametric linkage analysis
- By 2005/2006, clear that use of Illumina 4700 SNP panel offered equal if not superior information content
- As part of tender process after funding was obtained, an alternative possibility arose
  - To use Affymetrix 262,000 SNP chip
  - Turned out to be even cheaper than Illumina 4700 SNP option!

Genotyping platform

- For linkage analysis, 262,000 SNPs is overkill
  - One SNP every 0.5-1cM is more than sufficient
  - We ended up using a subset of 2,981 (LD-pruned, thinned to $\approx$ 1 SNP per cM) SNPs for linkage analysis
- But, allowed us to do a (mini) GWAS as well as linkage analysis
  - Using either family-based (TDT type) methods or case/control approach
    - Incorporating publicly available GWAS data from WTCCC controls
- However, final sample size (320 families, 694 cases, 136,310 SNPs following QC) not ideal for either...

Non-parametric linkage results (LOD>2)

SNP	Chr	cM	LOD	P
rs645490	2	32.903	2.227	0.007
rs9733150	11	87.935	2.032	0.0011
rs9977677	21	43.627	2.637	0.0002

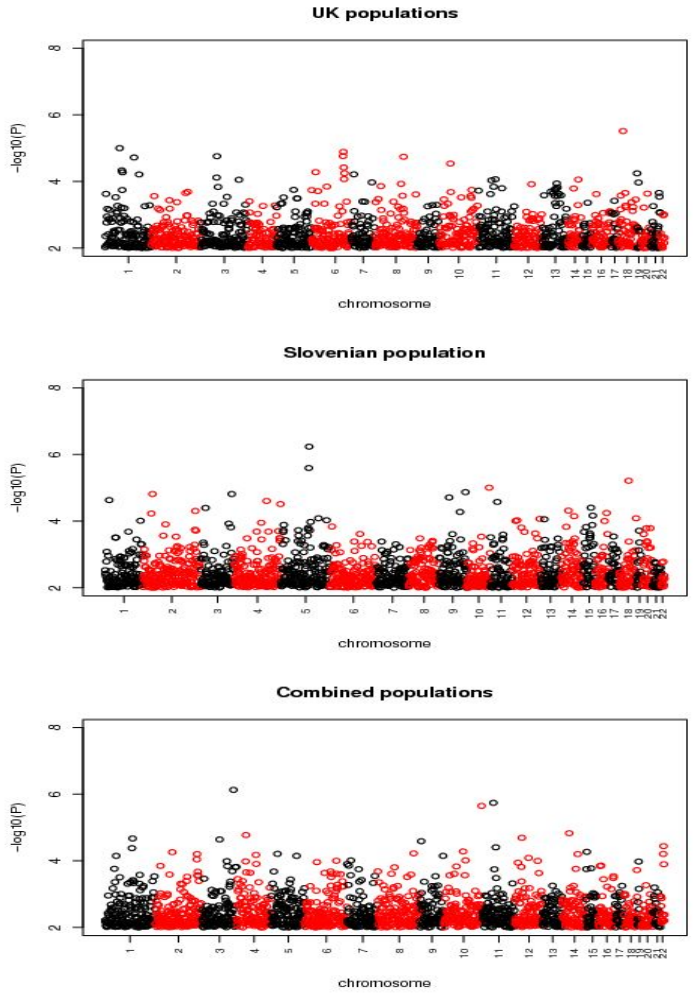
Parametric linkage results (HLOD>2)

	Closest non-parametric peak			Parametric Recessive					Parametric Dominant				
Chr	SNP	cM	LOD	SNP	cM	LOD	α	HLOD	SNP	cM	LOD	α	HLOD
2	rs645490	32.90	2.23	rs10184321	33.27	-11.98	0.23	2.17	rs645490	32.90	-31.76	0.28	2.61
3	-	-	-	rs484936	131.01	-13.42	0.18	3.02	-	-	-	-	-
10	rs1368532	154.66	1.55	-	-	-	-	-	rs1368532	154.66	-28.41	0.28	2.44
19	-	-	-	rs475188	92.61	-10.53	0.26	2.87	-	-	-	-	-
21	rs9977677	43.63	2.64	rs2835104	42.21	-8.61	0.28	3.21	rs9977677	43.63	-34.46	0.29	3.11

- Non-parametric peak on chr 10 increased in significance (LOD=2.32 at 160.38 cM, rs7904367) when using a narrower phenotype definition
  - VUR only, excluding RN without documented VUR
- Overall, little concordance between UK and Slovenian results
  - Or between our results and those of Feather et al. (2001)
  - Or with those of a Dublin group (Kelly et al. 2007)
    - Who performed a similar linkage scan in 129 families (283 affected individuals) of Irish descent, using 4710 SNPs

Association analysis results

- Nothing very compelling...

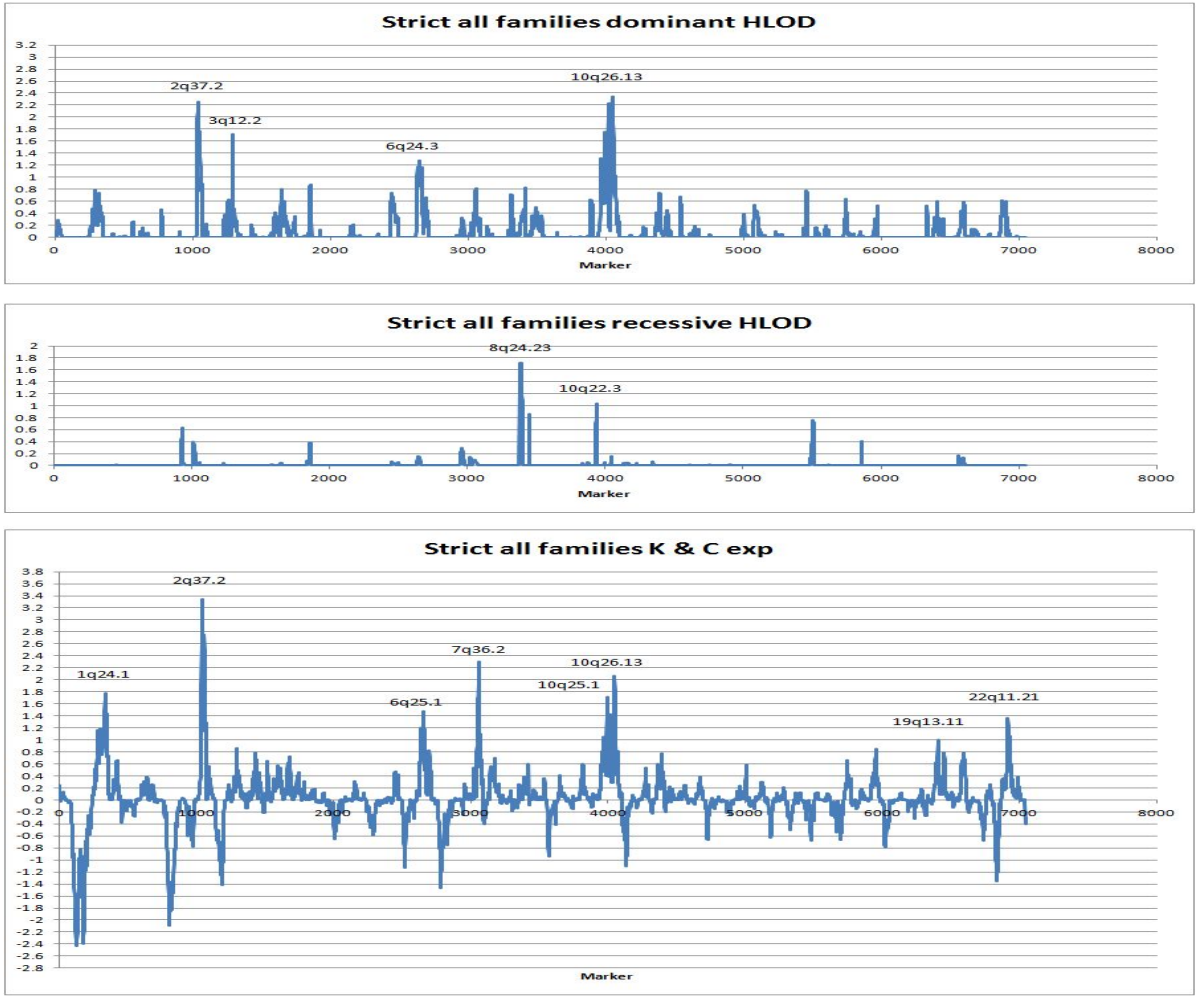


Updated Dublin scan

- In 2011 we embarked on a collaboration with the Dublin group to assist them with their analysis
  - Using an expanded set of $\approx$ 235 Irish families, typed on Affymetrix 6.0 array (643,691 post-QC SNPs)
  - Plus 851 Irish BioBank controls
- Association analysis (family-based or case/control) did not yield any compelling signals
- Linkage analysis was more promising
  - However, little concordance between results from 'old' and 'new' families
  - Or between combined results and previous studies (except perhaps on chromosome 10?)

Updated Dublin scan

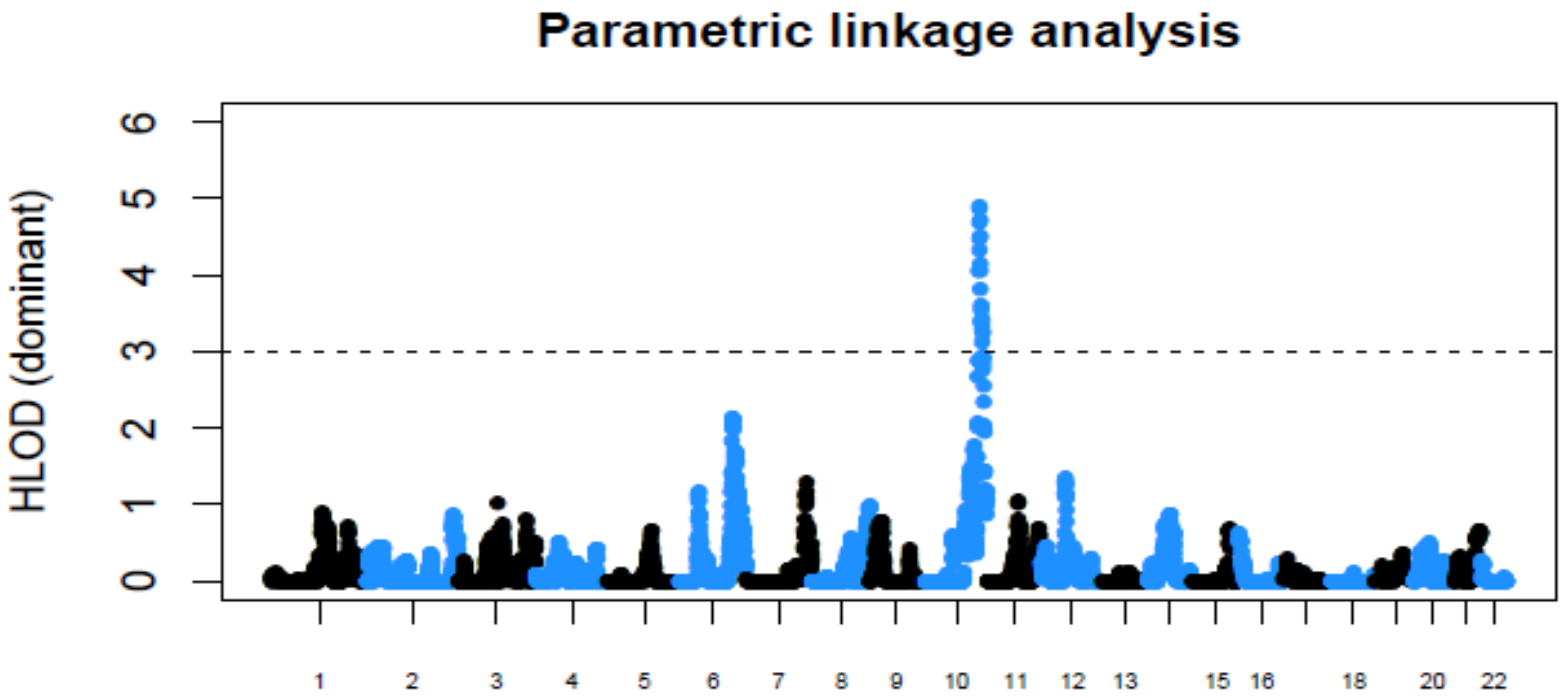
(Darlow et al. [2014] Mol Genet &Genomic Med)



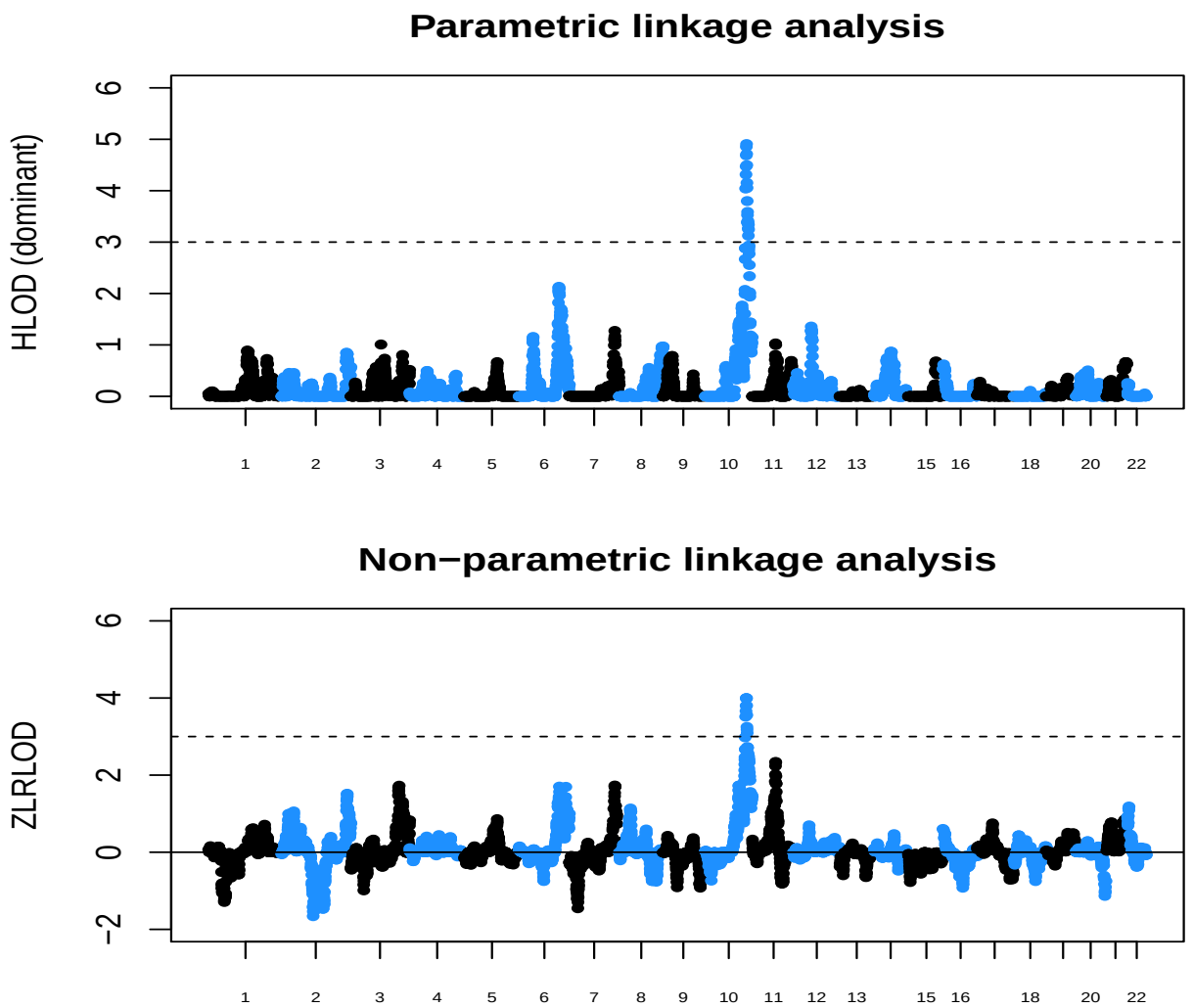
Combined analysis of UK/Slovenian/Irish families

- 548 families (460 informative for linkage)
  - Dublin: 235 families (500 affecteds)
  - UK: 165 families (303 affected offspring)
  - Slovenia: 148 families (353 affecteds, 313 affected offspring)
- SNPs available
  - 119,548 SNPs for family-based association analysis
  - 108,134 SNPs for case/control association analysis (with Irish BioBank and WTCCC controls)
  - 6922 (pruned and thinned to $\approx$ 2 SNPs per cM) SNPs for linkage analysis
- Association analysis (family-based or case/control) did not yield any compelling signals
- Nor did parametric linkage analysis under a recessive model
- However, with parametric linkage analysis under a dominant model...

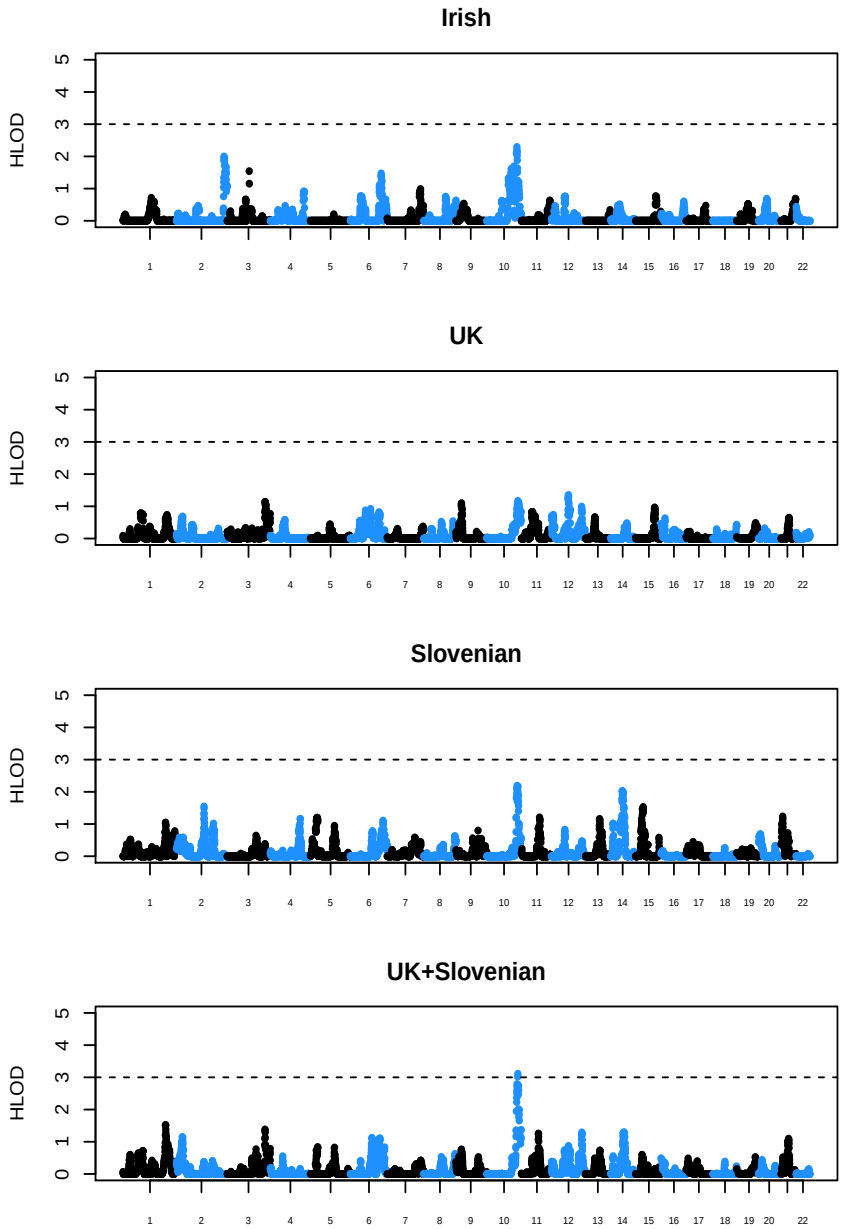
Combined (parametric) linkage results



Parametric vs non-parametric linkage results

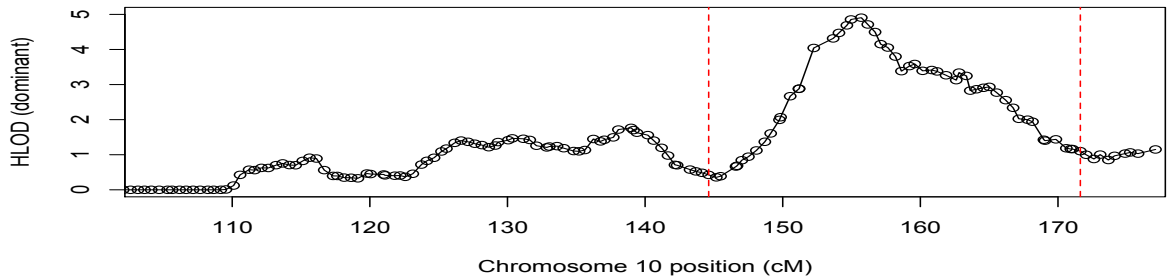


Individual studies

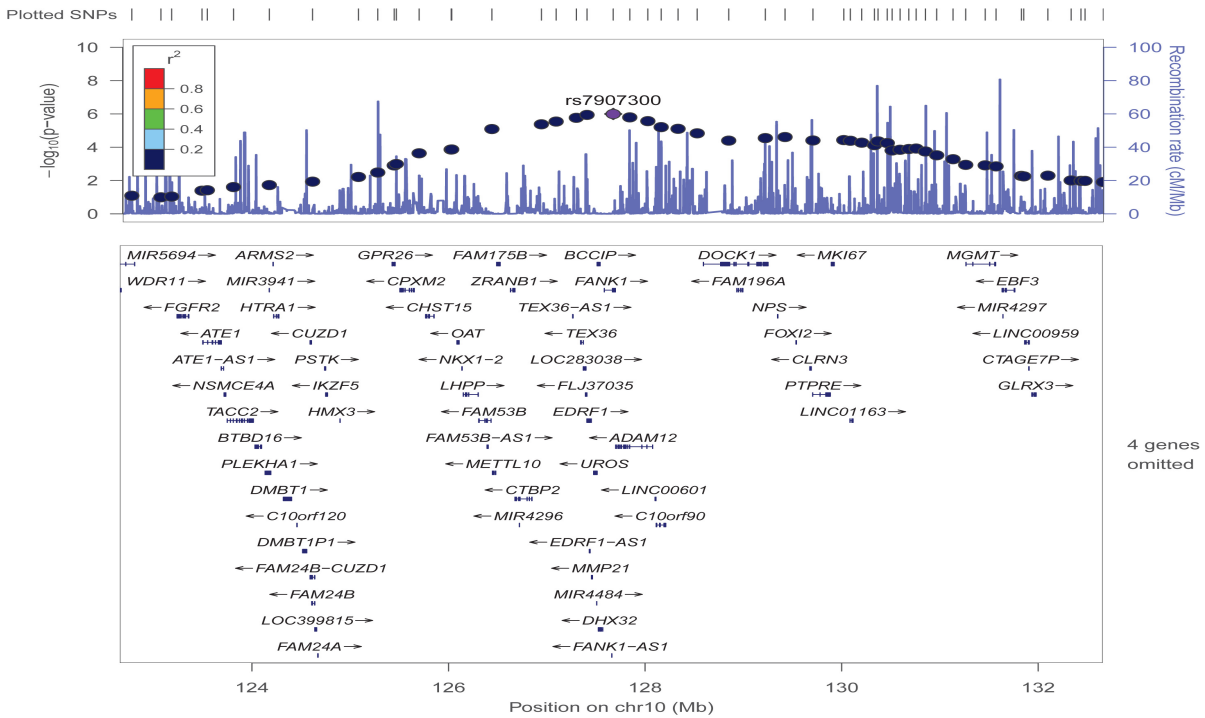


Chromosome 10 region

(a)



(b)



Follow-up of chromosome 10 region

- TDT association analysis of imputed SNPs in the linkage region did not produce any consistent/compelling signals
- Next step: sequencing under the linkage peak
  - To look for rare-ish variants in same gene (or same genomic window) that could explain the linkage signal
  - No requirement for the variants to be the same across families
    - Indeed from the HLOD analysis, we only expect $\approx 30\%$ of families to harbour mutations at the disease-causing genetic element in the 10q26 region
- We prioritized (larger) families showing the strongest evidence of linkage (two UK, two Slovenian, four Irish)
  - And whole-exome sequenced 29 individuals from these families (focussing our interest on the 10q26 region)

Whole-exome sequencing

- 37,930 non-synonymous variants (across whole exome) were called in 29 individuals
  - 139 lay within the 10q26 region of interest
  - Of these, 37 were relatively rare (MAF$<$ 0.05 in 1000G and Exome Sequencing Project)
  - Of these, 23 were found within same gene for two or more individuals (7 genes)
    - Six were predicted deleterious by one or more of SIFT/PolyPhen/ MutationTaster

Gene	Exonic variants	N predicted deleterious	N families
FANK1	5	2	3
CTBP2	6	0 (plus an additional 5 discounted as within segmental duplication)	6
MKI67	6	2 (plus an additional 1 discounted as within segmental duplication)	6
TACC2	2	0	2
DMBT1	2	2	3
ZRANB1	1	0	4
CLRN3	1	0	2

Table S4: Counts of rare variants (AAF < 0.05 in 1000GP/ESP) found in exome sequence of genes in the 10q26 region. Whole exome sequencing was carried out in 29 individuals from 8 families.

Whole-exome sequencing

- HOWEVER, only *FANK1* harbored rare variants that were actually consistent with the inheritance patterns/disease segregation/IBD sharing seen in multiple families

Position (GRCh37)	Variant (REF/ALT)	rsID	AAF (CEU)	Amino Acid change	family	SIFT score	Prediction
127668751	C/T	rs17153879	0.005	P>L	SLOV_028	0.03	Deleterious
127668863	GAA/-	rs146106149	0.025	E deletion	DUB_15	-	In-frame deletion
127693479	A/G	rs41302923	0.015	H>R	SLOV_028	0.02	Deleterious
127697001	G/A	rs139642438	0.001 (ESP6500)	G>E	SLOV_121	0.32	Tolerated
127697997	G/T	rs17153976	0.005	C>F	SLOV_028	0.45	Tolerated

Table S5: Exome (coding) variants found in *FANK1* in 3 families from whole exome sequencing. All are consistent with IBD sharing within each family.

- Investigation of *FANK1* using 189 European sequences from 1000G identified 715 variants, of which 407 were 'rare' (AAF < 0.05)
  - Only 7 sequences had no rare variants at all, suggesting that the detection of rare variants in *FANK1* in our VUR cases is not a rare event, and so may not necessarily be related to their phenotype.

Whole-exome sequencing

- HOWEVER, only *FANK1* harbored rare variants that were actually consistent with the inheritance patterns/disease segregation/IBD sharing seen in multiple families

Position (GRCh37)	Variant (REF/ALT)	rsID	AAF (CEU)	Amino Acid change	family	SIFT score	Prediction
127668751	C/T	rs17153879	0.005	P>L	SLOV_028	0.03	Deleterious
127668863	GAA/-	rs146106149	0.025	E deletion	DUB_15	-	In-frame deletion
127693479	A/G	rs41302923	0.015	H>R	SLOV_028	0.02	Deleterious
127697001	G/A	rs139642438	0.001 (ESP6500)	G>E	SLOV_121	0.32	Tolerated
127697997	G/T	rs17153976	0.005	C>F	SLOV_028	0.45	Tolerated

Table S5: Exome (coding) variants found in *FANK1* in 3 families from whole exome sequencing. All are consistent with IBD sharing within each family.

- Investigation of *FANK1* using 189 European sequences from 1000G identified 715 variants, of which 407 were 'rare' (AAF < 0.05)
  - Only 7 sequences had no rare variants at all, suggesting that the detection of rare variants in *FANK1* in our VUR cases is not a rare event, and so may not necessarily be related to their phenotype.
- Moreover the 'rare' *FANK1* variants listed above are actually much too common to realistically be considered as causes of VUR, without invoking a model involving extreme reduced penetrance

Targetted sequencing of 10q26 region

- No convincing causal variants had been found from exome-sequencing (in the 8 families examined)
- We therefore moved on to targetted sequencing of 10q26 linkage region
  - With an idea that there could be causal variants within non-coding (e.g. regulatory) regions
- We chose to sequence 32 unrelated affected individuals, from those families that contributed the most to the linkage signal
  - (the 8 families that had been exome-sequenced, and 24 others)

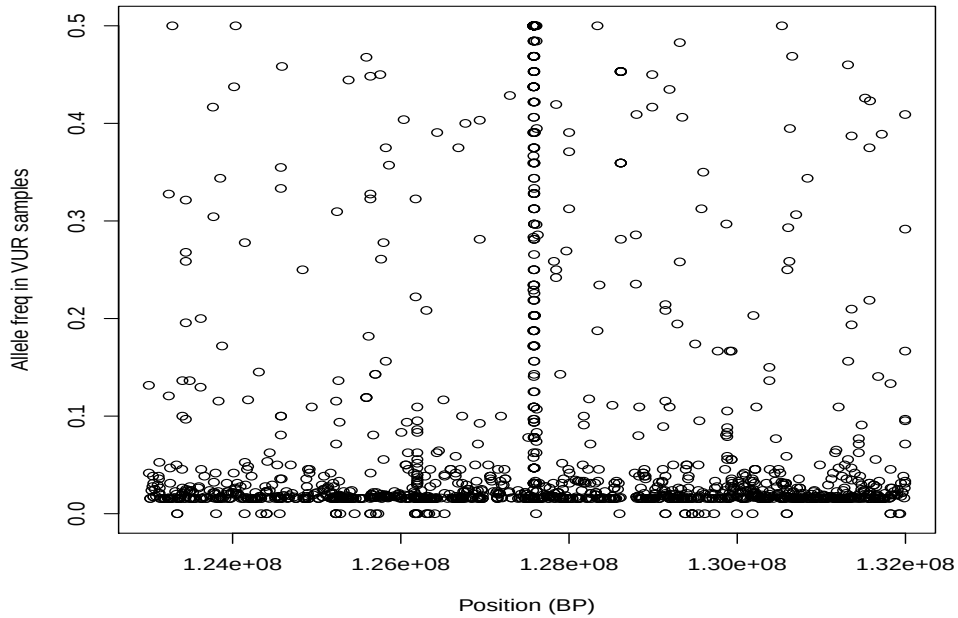
Targetted sequencing of 10q26 region

- No convincing causal variants had been found from exome-sequencing (in the 8 families examined)
- We therefore moved on to targetted sequencing of 10q26 linkage region
  - With an idea that there could be causal variants within non-coding (e.g. regulatory) regions
- We chose to sequence 32 unrelated affected individuals, from those families that contributed the most to the linkage signal
  - (the 8 families that had been exome-sequenced, and 24 others)

- 34,412 SNVs passed standard QC
  - Of these, 9170 were either absent or present at $< 1\%$ frequency in 1000G (Europeans)
  - Of these, 3625 were also shared between affected individuals (i.e. were present in at least 2 individuals sequenced)
    - $\Rightarrow$Sequencing artefacts?
    - Due to differences in read depth, alignment and variant-calling pipelines between our VUR cases and 1000G 'controls'?

Targetted sequencing of 10q26 region

- Focussing on variants not present in 1000G:



- Strong signal at *FANK1* – real or artefact?
- Concentrating on rare SNPs within promotors, we found 69 of 80 promotor-region SNVs occurred within the promotor of *FANK1*
  - Plus 11 occurring within the promoter of other genes

Rare SNPs within promotors

Position (GRCh37)	Variant (REF/ALT)	rsID	Gene	N individuals	AAF in VUR cases
124133947	G/C	rs72830779	PLEKHA1	2	0.03125
124220567	G/A	rs148152439	HTRA1	2	0.03125
126138911	G/A	rs140024768	NKX1-2	2	0.03125
129691655	A/G	rs182013372	CLRN3	2	0.03125
129924568	T/A	rs151328961	MKI67	3	0.09375
131934190	T/C	rs11017099	GLRX3	8	0.125
131934191	A/G	rs11017100	GLRX3	8	0.125
131934201	A/G	rs11017101	GLRX3	8	0.125
131934203	C/T	-	GLRX3	8	0.125
131934207	T/G	-	GLRX3	8	0.125
131934212	T/C	rs11017102	GLRX3	8	0.125

Table S6: Rare variants (AAF<0.01) shared in 2 or more individuals, found in the promoter region of genes other than *FANK1*, from targeted sequencing of 10q26.

- Of the non-*FANK1* genes, arguably most convincing is *GLRX3*
  - Harbors a rare haplotype seen in 8 out of 32 VUR cases
  - However, not a good positional candidate
    - (lies at far end of linkage region)
  - Some additional cause for concern from inspection of sequencing reads
    - Alternate alleles only seen in first or last 25bp of any given read
    - And only in $\approx 15\%$ of high-quality reads
  - $\Rightarrow$ Requires validation

FANK1

- Further interrogation of the .bam files showed that the *FANK1* region appears hypervariable
  - Many of the called SNVs were clearly artefacts
- 69 of the *FANK1* SNVs appeared potentially genuine
  - Including a G/C SNV at 127,584,687 bp, and an adjacent C/A SNV at 127,584,688 bp
    - Both were heterozygous in all 30 (out of the 32) individuals in which they was called
    - However we only expect putative causal variants to be shared in around 10 to 15 of our samples – the sharing of the same causal variant by 30 affected individuals from 30 different families seemed somewhat suspicious...

FANK1

- Our suspicions were confirmed when we performed a BLAST search on the 500 bp sequence around these SNVs
  - The sequence containing the alternative (C and A) alleles at 127,584,687 and 127,584,688 bp showed high levels of sequence similarity to the short arm of chromosome 22 on the GRCh38 genome build
    - This region is unmapped in GRCh37
- We re-aligned our targeted sequences to the GRCh38 assembly
  - 9 of the 11 previously identified SNVs in the *FANK1* promoter region were found to align fully to chromosome 22
    - Only the two at 127,584,687–127,584,688 bp (125,896,118–125,896,119 bp on GRCh38) mapped to chromosome 10
  - Suggests that a number of the reads originally identified as mapping to 10q26 may really come from chromosome 22
  - Offers a possible explanation for the unusually high level of variability observed...

FANK1

- Further investigation revealed that the 42.4 kb region of chromosome 10 from 125,885,884 to 125,928,375 contains segmental duplications from at least 4 other genomic locations
- These segmental duplications are most likely to have generated the high proportion of rare variants recorded for *FANK1* and its promoter
  - Which therefore represent artefactual findings unrelated to the VUR phenotype...
- More sophisticated experimental methods (such as amplification via long-range PCR prior to sequencing, or the use of longer-read sequencing technologies) will be required in order to robustly interrogate this 42.4 kb region, without danger of contamination from other genomic regions

TASER analysis

- Two papers have addressed the issue of comparing cases and controls that were sequenced separately (at different depths)
  - Derkach et al. (2014) "Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic" (Bioinformatics 30:2179-88)
  - Hu et al. (2016) "Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls" (PLoS Genetics 12:e1006040)

TASER analysis

- Two papers have addressed the issue of comparing cases and controls that were sequenced separately (at different depths)
  - Derkach et al. (2014) "Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic" (Bioinformatics 30:2179-88)
  - Hu et al. (2016) "Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls" (PLoS Genetics 12:e1006040)
- We obtained the .bam files from 1106 ALSPAC controls sequenced as part of the UK10K study
  - Low-coverage (2–20 fold)
  - Compared to up to 1000 fold coverage (in the target region) for our 32 VUR cases
- We used TASER to effectively re-call variants in both cases and controls, allowing for systematic differences in read depth (and other factors), at the same time as constructing a test statistic

TASER analysis

- TASER splits the entire genome into small windows of interest (e.g. genes or exons of genes)
  - TASER then uses the total number of reads mapped to a variant, and the number carrying the minor allele, to calculate a score statistic at each position in a gene (or window) of interest
    - Provides an assessment of the effect of each individual variant on the disease phenotype

TASER analysis

- TASER splits the entire genome into small windows of interest (e.g. genes or exons of genes)
  - TASER then uses the total number of reads mapped to a variant, and the number carrying the minor allele, to calculate a score statistic at each position in a gene (or window) of interest
    - Provides an assessment of the effect of each individual variant on the disease phenotype
- A burden statistic is then calculated for each window as the sum of the score statistics for each of the variants within that window
  - Allows identification of windows that have a higher or lower accumulation of rare variants in the cases than might be expected, compared to controls
    - Effectively looks for windows containing clusters of rare variants seen in different affected individuals (and not seen - or at least not at comparable frequencies - in controls)

TASER analysis

- TASER splits the entire genome into small windows of interest (e.g. genes or exons of genes)
  - TASER then uses the total number of reads mapped to a variant, and the number carrying the minor allele, to calculate a score statistic at each position in a gene (or window) of interest
    - Provides an assessment of the effect of each individual variant on the disease phenotype
- A burden statistic is then calculated for each window as the sum of the score statistics for each of the variants within that window
  - Allows identification of windows that have a higher or lower accumulation of rare variants in the cases than might be expected, compared to controls
    - Effectively looks for windows containing clusters of rare variants seen in different affected individuals (and not seen - or at least not at comparable frequencies - in controls)
- We split our target region into consecutive 120 bp windows
- Only bases called with a quality score > 30 were added to the read count at each position

TASER results

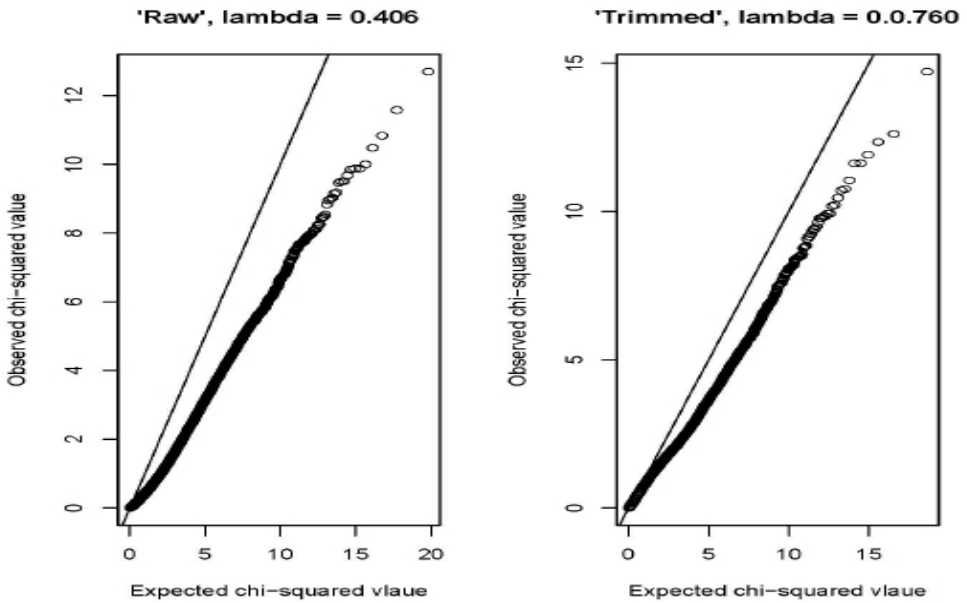


Figure S10: Quantile-Quantile (Q-Q) plots showing the ordered -log10(p-values) from TASER against their expected values under the null hypothesis of no differences between cases and controls. Left hand plot shows the 'raw' results where every allele seen in any sequencing read is counted. Right hand plot shows the 'trimmed' results where only alleles seen at least 5 times in that position in the cases (high read depth) and at least twice in the controls (low read depth) are counted.

- (Only counting alleles that are seen at least twice in the low depth control sequences, and at least 5 times in the high depth VUR sequences, – in order to reduce the expected number of sequencing artefacts – improved the overall distribution of test statistics)

Top TASER results

Window (bp, GRCh38)	L	M_st	M_p	STB_p	Probable variant position	Variant (REF/ALT)	rsID	VUR AAF (n= 32)	ALSPAC AAF (n=1106)	Gene
124841200-124841319	1	1	1	6.52x10-4	124841267	C/A	rs533072490	0.04	0.0	3 kb upstream FAM175B
125655040-125655159	1	1	1	4.43x10-4	125655088	A/T	rs183471950	0.0468	0.0016	TEX36
125679640-125679759	1	1	1	3.83x10-4	125679648	G/A	rs541040960	0.0313	0.0	TEX36
127059640-127059759	1	1	1	1.25x10-4	127059696	A/AT	rs551248465	0.0468	5.7x10-4	DOCK1
128329480-128329599	1	1	1	5.58x10-4	128329528	A/G	rs146303503	0.0468	0.0012	10 kb downstream LINC01163
128770720-128770839	1	1	1	8.88x10-4	128770839	G/A	-	0.0313	0	400 kb downstream LINC01163
128929840-128929959	1	1	1	6.51x10-4	128929870	G/A	rs182100352	0.0313	0	500 kb upstream MGMT

Table S8: The 7 out of the top 10 TASER results that were considered reliable, with MAF threshold 0.05 where L = number of variants "seen" in a particular 120 bp window; M_st = number of variants passing screening and threshold procedures; M_p = probable "true" variants; New-STB_p = Burden statistic p value. VUR and ALSPAC MAFs calculated from read counts at the variant position as determined by inspection of the appropriate "read-count" input file.

Conclusions

- Parametric (and non-parametric) linkage analysis of 548 ASP families identifies a region on chromosome 10q26 with strong evidence of harboring genetic variants contributing to VUR
  - Identifying the causal variant(s) underlying the linkage signal has proved more challenging
  - A 42.4 kb region (containing the key candidate gene *FANK1*) remains effectively untargeted by our sequencing endeavours, on account of containing segmental duplications from other genomic locations
    - Will need to do long-range PCR prior to sequencing, or use longer-read sequencing technologies, to robustly interrogate this region

Conclusions

- Parametric (and non-parametric) linkage analysis of 548 ASP families identifies a region on chromosome 10q26 with strong evidence of harboring genetic variants contributing to VUR
  - Identifying the causal variant(s) underlying the linkage signal has proved more challenging
  - A 42.4 kb region (containing the key candidate gene *FANK1*) remains effectively untargeted by our sequencing endeavours, on account of containing segmental duplications from other genomic locations
    - Will need to do long-range PCR prior to sequencing, or use longer-read sequencing technologies, to robustly interrogate this region
- Variant calls from next-gen sequencing do NOT represent 'truth'!
  - Highly dependent on the sequencing platorm and bioinformatics platforms used
  - Potentially full of artefacts...

Conclusions

- Parametric (and non-parametric) linkage analysis of 548 ASP families identifies a region on chromosome 10q26 with strong evidence of harboring genetic variants contributing to VUR
  - Identifying the causal variant(s) underlying the linkage signal has proved more challenging
  - A 42.4 kb region (containing the key candidate gene *FANK1*) remains effectively untargeted by our sequencing endeavours, on account of containing segmental duplications from other genomic locations
    - Will need to do long-range PCR prior to sequencing, or use longer-read sequencing technologies, to robustly interrogate this region
- Variant calls from next-gen sequencing do NOT represent 'truth'!
  - Highly dependent on the sequencing platorm and bioinformatics platforms used
  - Potentially full of artefacts...
- Use of tools like TASER can help improve inference when these artefacts are likely to vary between cases and controls
  - Sequencing (a subset of) controls along with the cases is highly recommended

Acknowledgements

SCIENTIFIC REPORTS

OPEN

Received: 6 July 2017
Accepted: 6 October 2017
Published online: 03 November 2017

Genome-wide linkage and association study implicates the 10q26 region as a major genetic contributor to primary nonsyndromic vesicoureteric reflux

John M. Darlow[1,2], Rebecca Darlay[3], Mark G. Dobson[1,2], Aisling Stewart[3], Pimphen Charoen[4,5], Jennifer Southgate[6], Simon C. Baker[6], Yaobo Xu[3], Manuela Hunziker[2,7], Heather J. Lambert[8], Andrew J. Green[1,9], Mauro Santibanez-Koref[3], John A. Sayer[3], Timothy H. J. Goodship[3], Prem Puri[2,10], Adrian S. Woolf[11,12], Rajko B. Kenda[13], David E. Barton[1,9] & Heather J. Cordell[3]

- Rebecca Darlay, John Darlow

- Wellcome Trust
- Medical Research Council